Identification of Emerging Scientific Topics in Bibliometric Databases

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Inform.Wirt Carolin Mund

Tag der mündlichen Prüfung:	.....
Referent:	Prof. Dr. Rudi Studer
Korreferent:	Prof. Dr. Ulrich Schmoch

Tag der mündlichen Prüfung 18.07.2014

One is always considered mad,
when one discovers something
that others cannot grasp.

Edward D. Wood, Jr. (1922-1978)

I’m deeply grateful to all people who contributed to the achievement of this thesis. You all found your personal way to help me out in times of need or to challenge me when I was getting too comfortable. The following short list can only hint at all the support I got, but it should give you an idea of how thankful I am:

My family and friends, Rudi, Uli, Rainer, Achim, my colleagues and friends from NISTEP and Fraunhofer ISI, …

Short Table of Content

I Overview

1 Introduction

This thesis is concerned with the assessment of novel methods to discover emerging topics in science. This chapter explores the motivation behind this task. In particular, it explains why such an approach is necessary (Section 1.1) and its essential characteristics (Section 1.2). The chapter ends with an outline of the thesis and its relation to the author's previous publications.

1.1 Motivation

This thesis addresses the identification of emerging topics in science. Typically, publication data are used to map progress in science and thus the emergence of novel topics. The goal of this thesis is to use these so-called bibliometric data to monitor the scientific landscape to detect emerging topics. The limitations of the data and the derivable indicators[1] have to be acknowledged in this regard. Therefore, a clear focus of this thesis is on the distinction between reliable and unreliable indicators of emerging topics. The necessity for indicators that are independent of citation or impact measures is illustrated by a publication by Mendel (1865), which is discussed in more detail during the course of this thesis and also serves as a running example:

Mendel wrote a highly innovative paper in 1865 titled "Versuche über Pflanzen-Hybriden (Experiments with Plant Hybrids)" (Mendel 1865). This paper represented groundbreaking work for the understanding of genetics and inheritance and acknowledgement and follow-up studies by the scientific community could have been expected regarding the findings. On the contrary, however, there were only two noteworthy reactions: A few researchers questioned his figures, but the majority ignored his work for decades (see e.g. Atkins 2003, pp. 46f, van Raan 2004). Only one "misleading" citation was made in 1881 (Atkins 2003, p. 47). 35 years later, Mendel's findings were confirmed or, more precisely, rediscovered (by Hugo de Vries, Carl Correns and Erich Tschermak) and only then acknowledged by the scientific community for the first time.[2] Similarly, but in a different field, the later findings by Planck were first "met by silence [... and] regarded as a mathematical ruse" (Atkins 2003, p. 205).

Reasons for the belated acknowledgement could be that the publication by Mendel "drowned" in the vast sea of scientific publications. It is true that, at that time, scientific publications were not produced in the same quantity as nowadays (for a discussion of growth rates in science, see Michels and Schmoch 2012), but the access to publication data was also unstructured and complicated. While the introduction of the internet has increased awareness of worldwide publications (and also facilitated

---

1 If not stated otherwise, the term indicator will denote any part of the system that enables the flagging of documents or topics (cluster) as "emerging" (or - based on the lack thereof - as "not emerging"). Indicators are calculated based on certain characteristics of the publications. Those characteristics that are computable, comparable and stable are labelled features and only these can form the basis for indicators. In themselves, features have no explanatory power about the emergence of a topic. However, indicators are generated by applying these features to rules, topic models etc. The concept of features is explained in more detail in Chapter 4.

2 http://www16.us.archive.org/stream/planthybridizati00robe/planthybridizati00robe_djvu.txt, last accessed on 2014/02/14.

open access – an option that was simply not possible with former publication means), past publications were restricted to physical outlets and thus also locally bound. Garfield (1970) argues that Mendel's paper would have been cited if the ISI Science Citation Index[3] had been around at that time: "I like to think that SCI will not only prevent inadvertent neglect of useful work but, feel confident it will prevent much unwitting duplication of research and publication" (Garfield 1970, p. 70).

Besides the restrictions related to the publication form, there are other known factors that might influence the reception of a publication. Even if a paper is read by a wide variety of researchers, in the end, its reusability in other (related) work and applications determines its dissemination and also the upper limit for its citation count. As Mendel's work was deemed useful in retrospect, other reasons must have prevented its recognition. One possible explanation is that Mendel's paper was refuted simply on the grounds of its high innovativeness, i.e. scientists were overwhelmed by the novelty of the findings. The gap between the state of knowledge prior to and after his groundbreaking work might have hindered other scientists relating it to their work (cf. Grinnell 1987, pp. 45f). However, another factor that will be discussed later in more detail is that Mendel had to rely on Mathematics to explain his findings – a fact that was not well received in his scientific environment (Barber 1961, Atkins 2003, p. 47). Regardless of the exact or main reason, in the end, the relevance of Mendel's work was acknowledged, albeit belatedly.

Given this background, it is important to grasp how humanity evolves as an intelligent species; discoveries and errors (and errors are merely "negative" discoveries"[4]) are passed on to other humans and generations (cf. Section 2.1, Johnson 2013, p. 172). This spread of knowledge avoids repetition of efforts, errors and failures and ensures the advancement of science at the research front – the point of development where humanity is currently positioned – in contrast to the knowledge level of individuals or groups.

Thus, it becomes irrelevant whether these earlier discoveries were made by the same person, group, country etc. Regardless of their source, they form the basis for further common advancement. Nonetheless, as the above example and later discussions show, the selection of related work can be biased due to various factors (see in particular Sections 2.2 and 3.2.1). However, in the ideal case, research is based on the most recent discoveries in a scientific field.

By developing a system for the semi-automatic detection of emerging topics in science, this thesis aims to facilitate their accessibility and observance. The resulting procedure is comparable to detecting outliers in a set of documents. The main problem with outlier detection is the correct definition of the selection criteria. In this thesis, different indicators were therefore tested for their applicability. All of them were calculated using bibliometric data available at the time of publication.

---

3 See explanation of Web of Science below, Section 3.1.

4 Cf. with the "most famous failed experiment in history" by Michelson and Motley, which built the foundation for the refutation of the aether theory (Blum and Lototsky 2006, p. 98).

Independent of their application in the resulting system, these indicators can also be interpreted independently to enable a better understanding of the evolution process of emerging topics. In particular, drivers of and factors for the success of new scientific topics can be derived.

Deployed in the system developed in this thesis, the indicators enable the detection of publications in novel scientific topics and the extraction of deviant publications. The selected publications are presented to an end user for further inspection. This presentation not only serves as a signal but also facilitates the monitoring of ongoing research.

Based on the publication set presented to him, the end user can derive new topics. The focus on single documents takes into account their role as catalysts for emerging topics. Every milestone in science, be it a discovery, a publication etc., starts as a line of thought. Some work aligns with previous developments but creates novel ideas so that the line of thought branches. In this sense, earlier research is merely the clay that is used as the input for new experiments, discoveries or thoughts. And similar to using clay, regardless of its colour or consistency, old ideas from different backgrounds are mashed up to form new developments that were never intended by their original producers.

The resulting "paradigm", as defined by Kuhn (1973, pp. 10f), is a unique and new topic in the scientific landscape that still leaves enough room and is engaging enough for further scientific activity.[5] As such, a so-called "paradigm shift" establishes new foundations and frameworks for scientific endeavours. The incorporation of discoveries, new findings and new topics in the scientific landscape thus opens up new opportunities for research and allows the advancement of science as a whole.

When imagining scientific progress as a simple linear equation, a new discovery might be just one more input parameter, e.g. one experiment, away. The publication of discoveries creates new input parameters from former outputs and allows the aforementioned local and organizational independent usage of knowledge. The underlying assumption is that given enough time and findings (by oneself or other people) as input parameters, at some point, a new discovery is the output. Thus, the knowledge creation process becomes a self-enhancing cycle. In this case, the institutionalization of paradigms assures that research is conducted efficiently. In the scientific community, this can be seen as a globally spanning knowledge cycle similar to that of a knowledge management system within an organisation (cf. Staab et al. 2001). The scientific publication process allows the cycle to be elevated to a global level:

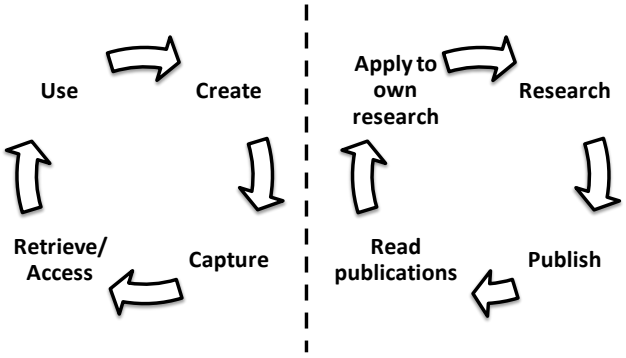
- Knowledge creation: A "knowledge worker", in this case a scientist, makes a discovery
- Capture: The so created knowledge is recorded in a scientific publication
- Retrieve/Access: The publication can be accessed by the scientific community
- Use: The knowledge is applied in new contexts or for further work on that topic[6]

---

5 "[Such work is] sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity [...but also] sufficiently open-ended to leave all sorts of problems for the practitioners to resolve." (Kuhn 1973, pp. 10f).

6 This is the same as the original definition by Staab et al. (2001, p. 32): "the knowledge worker will not only recall knowledge items, but she will process it for further use in her context."

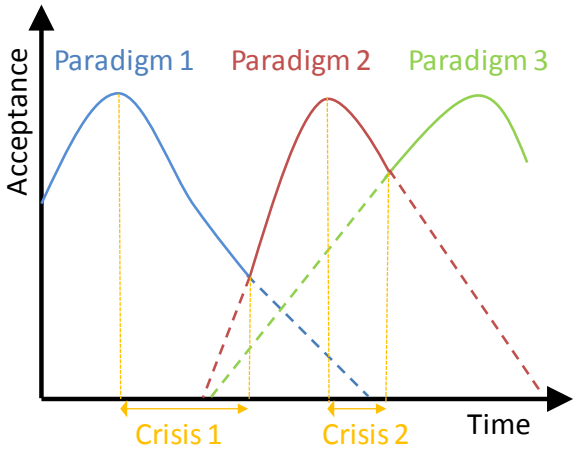
Figure 1: The knowledge cycle as described by Staab et al. (2001, p. 27) shown on the left hand side, and transferred to the scientific communication process on the right.



Source: Own illustration, partly adapted from Staab et al. (2001, p. 27)

Figure 1 compares the initial knowledge cycle as given by Staab et al. (2001) with its described transfer to the scientific communication process. “Research” as an action represents a researcher working at his desk, in the lab, at the test bed etc.

Figure 2: The interplay between the acceptance of a paradigm and its development.



Source: Own illustration, idea from Kuhn (1973, pp. 64f, 67f, 71, 77)

While the knowledge cycle works on a small scale for individuals (or even as product cycles), the scientific evolution process can be described analogously to cycles in economics on a large scale (like Kondratieff waves). According to Kuhn (1973, pp. 62ff), the drivers behind scientific revolutions are “paradigm shifts”. A paradigm in Kuhn’s sense can be any establishment of concepts, rules or methods as the foundation of ongoing research. Thus, a paradigm shift represents the replacement of outdated beliefs on the basis of new findings. Acceptance plays a major role in this context and is a recurring aspect in this thesis.

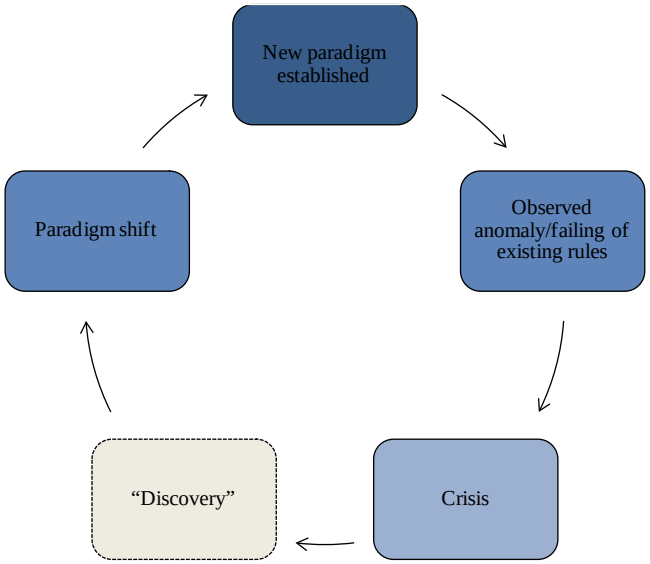
Figure 2 illustrates the relationship between the acceptance of a paradigm and its substitution. Kuhn refers to acceptance indirectly when writing about paradigm shifts: A paradigm is neither questioned nor replaced as long as no “anomaly” is encountered (Kuhn 1973, p. 62). An anomaly is a contradic-

tion to the rules stated by the paradigm.[7] Depending on the severity of the anomaly and its recurrence, a "crisis" can be evoked. This crisis must not necessarily be recognized as such, but it stimulates the work on alternative ideas and rules and thus the "discovery" of new paradigms (Kuhn 1973, p. 86). In fact, a crisis is not a necessary precondition of a paradigm shift, but rather a catalyst for one.

In any case, at some point, a new paradigm is found that resolves the anomalies posed by the current one. Figure 2 demonstrates that, during a crisis, the acceptance of a paradigm declines constantly, but due to the lack of alternatives, the scientific community still clings to it. If, and only if a new paradigm has been established, is the old scientific theory rejected (Kuhn 1973, p. 77). In time, the acceptance for the new paradigm increases. In the words of Abernethy and Sparrow (1992, pp. 12f), the two paradigms "battle for acceptance".

Again, Kuhn indirectly mentions the importance of acceptance in the scientific community during a paradigm shift: "[...] novelty emerges only with difficulty, manifested by resistance, against a background provided by expectation" (Kuhn 1973, p. 64). By definition, a paradigm shift as defined by Kuhn can only be an improvement to the scientific status quo; since the concepts of the old (to be replaced) paradigm are taken as a given, the new paradigm must have a higher explanatory power, applicability and thus utility. And if this is the case, a paradigm prevails over its competitors (Kuhn 1973, p. 23). Thus, what it boils down to in the end is the acceptance of a paradigm in the scientific community. In this context, it should be also noted that progress, e.g. in the sense of crossing boundaries for new developments, is often triggered by the activities of single researchers (cf. Thompson Klein 1996, pp. 35, 44) – as was the case in the running example of Mendel. Thus, the acceptance of ideas or paradigms is closely linked to the acceptance of individuals (cf. Finger 2000, pp. 305f, Fang 2014, Grinnell 1987, pp. 49f).

Figure 3: The cycle of constant paradigm shift in science.



Source: Own illustration, idea from Kuhn (1973, pp. 64f, 67f, 71, 77)

7 In a similar notion, Fang (2014) shows how discrepancy between the "suitable scope" and the "examined scope" can lead to a crisis.

Figure 3 illustrates the paradigm shift as described above. Here, the concept of a discovery is introduced, which triggers the paradigm shift. However, the image of such a discovery as a scientist experiencing a "eureka" moment in his laboratory, under an apple tree etc. is rather unrealistic. "Discovery" is an artificial construct to instantiate a theoretical or conceptual idea. In a succession of events, the result of scientific progress is thus "attributed to an individual and to a moment in time" (Kuhn 1973, p. 55) to represent the "consequence of the whole process through which change is analyzed, debated and assessed" (Schaffer 1994, p. 19). Ultimately, the respective persons and actions are chosen by the scientific community depending on their impact in terms of applicability, utility etc. (Schaffer 1994, pp. 13f). Thus, again, an action can only be denoted a discovery as long as its circumstances are accepted by fellow researchers. In Mendel's case, this led to a time lag between the event and its labelling as a discovery (cf. Schaffer 1994, pp. 13f). In turn, the act of singling an action out as a discovery facilitates the dissemination and acceptance of a new finding.

In Kuhn's sense, the discovery, i.e. the active and conceptual work it represents, is the information that was missing in the old paradigm, which fills the gap that is (repeatedly) red-flagged by anomalies during the crisis (Kuhn 1973, pp. 77f). Progress in science thus depends on the dissemination of information or knowledge and its acknowledgement by the scientific community. In the course of this thesis, a closer look will therefore be taken at the diffusion and acceptance of novel topics in science and, primarily, a system developed that spotlights emerging topics.

Such a system could help to disseminate scientific findings, which – similar to economic goods – compete for consumers' attention. Like products, knowledge that is not used by anyone is worthless. Propaganda is thus important and is typically achieved via publishing. The phrase "publish or perish" not only applies to researchers, but indirectly to ideas: ***be*** published or perished. Publication is only the first step on the producer's side but an idea may perish if it is not acknowledged, as was the case with Mendel. It seems natural that some ideas are just not fit enough to survive and have to make room for other (better) ideas (cf. van Dalen and Klamer 2005, p. 399). This applies to both the evolutionary and economic counterparts. Thus, "publicity" is one key factor for the success of a scientific publication.

Even though access to publications is much easier today, the fact that the main part of citation research is merely "browsing" makes the process and thus the outcome nondeterministic. This is a natural consequence as the process is a non task-oriented, non system-oriented document study (Vakkari 1999). Since the researcher "does not have a precise objective in mind [... he] is simply exploring information pathways that appear to be exciting and interesting" (Garfield 1984, p. 530). Modern search engines and databases should enhance the chance of a (serendipitous) retrieval of sought-after knowledge (cf. Section 2.1), but new findings are frequently "overlooked", even if they have already triggered a trend in a sub-group of scientists. Although most of the reasons for this are to do with human nature, they could still be eradicated with the help of a method that systematically points out the blind spot of the current procedure (cf. Garfield 1970 as referred to above).

The importance of a "signalling effect" should be highlighted, especially for new findings and emerging topics: In an early development stage, no concept is known to "outsiders". Because of this and other factors that will be discussed during the course of this thesis (see in particular Sections 2.2 and 2.3 and Chapter 8), they are easily "overlooked" by the scientific community. A search explicitly for

emerging topics becomes necessary.[8] Famous examples of innovative findings that forged new paths in science include the discovery of the double helix, the already discussed findings in genetics by Mendel or the Higgs boson (Higgs 1964, Englert and Brout 1964, O'Luanaigh 2013). As the examples show, the scientific community might not (immediately) recognize the importance of the new topics.

Furthermore, new topics are not necessarily characterized by interdisciplinarity. However, they all enhance an existing topic by adding a new perspective, problem or solution. This enables scientists to work on issues that were not possible before.[9] It might be the case that a theory or a discovery is proven wrong later or improved in a later iteration, but this does not diminish the level of potential and innovativeness of the topic at the point in time when it was discovered. As was already hinted at before, the greatest advancements in science are made thanks to the combination of different points of view, approaches etc. which in turn demands personalities from different backgrounds (cf. Chapter 7).

These discoveries are important – in themselves and regardless of their later fate – as they provide new "food for thought"; they avoid stagnation in science because – as mentioned earlier – even the pursuit of false paths in combination with later rectification is progress. If the associated experience enters the knowledge cycle as described above, it can be used in the creation of new knowledge. This goes hand in hand with the view that a new paradigm is always an improvement (cf. Kuhn 1973, pp. 64f).

Therefore, this thesis tries to neutrally identify emerging topics. Their importance or impact can be decided by an expert or by the scientific community later. In particular in the latter case, the fate of a topic depends on many outside factors that should play no part in its assessment. Funding programs that target risky endeavours like topics with unusual or unstable settings and an explorative nature are also emerging to support the knowledge creation process at an early stage and avoid unnecessary dead-ends.[10] This political interest in new topics in science shows that monitoring ongoing research is in fact more important than or at least independent of their actual success.

The difference between a new topic and a new idea is the scope of impact. With the introduction of a new topic, the overall setting in which the subsequent publications were framed changes completely. Previous work might no longer be applicable in this case. Either a) it does not cover the subject sufficiently or b) the new discovery changes the view tremendously. The former might be the case if for instance the problem arises once two previously separate scientific areas are combined. Findings from one area were not included in the other area's research (cf. exaptation in Section 2.1). Now, the combined efforts may shed new light on former unexplained observations or "rejuvenate the merged components" (Swanson 1993, p. 606).[11]

---

8 In the course of this thesis, different definitions of new ideas or new scientific topics will be discussed (cf. Section 2.1). For now, it is sufficient to say that a new topic emerges on the basis of a discovery or similar aspects.

9 Cf. discussion of paradigm shifts as defined by Kuhn (1973, p. 12).

10 See for instance http://ec.europa.eu/programmes/horizon2020/en/h2020-section/future-and-emerging-technologies, last accessed on 2014/05/27.

11 For instance, combining elements of Mathematics and Biology formed the foundation for Genetics in Mendel's case.

With the help of a system that increases awareness of new topics at an early stage, new findings could be included in ongoing research with a shorter delay. Without it, unnecessary and pointless efforts or the reinvention of the wheel can be the consequence. This is inefficient and a waste of time and resources. Note that previous findings might be unintentionally ignored. In this case, some kind of signal could help to avoid this mistake. Since topics stagnate as long as they arouse no interest, constant awareness of new findings or topics is important to keep research at the cutting edge. The research publication process as a global knowledge management tool thus forms the backbone of the scientific knowledge cycle. It can be improved by introducing new implementations of its individual components. In this regard, this thesis delivers a tool for the retrieval/access part of the knowledge cycle. In the following section, the requirements of such a system are discussed.

1.2 Problem Statement

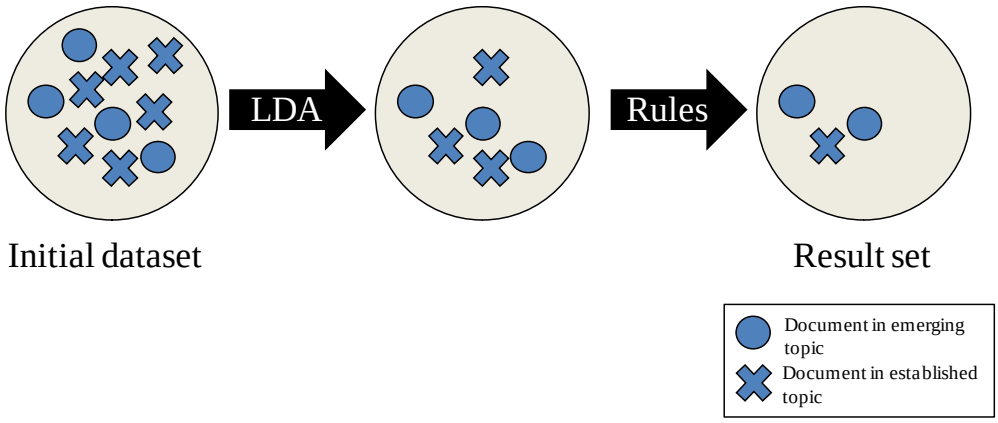
Section 1.1 explained the close link between acceptance and dissemination of new findings and the consequent need for heightened awareness of them. In the course of this thesis, other factors influencing the recognition of emerging topics will be discussed. The understanding and acceptance of methods play a crucial role here. In our example, the reasons for the rejection of Mendel's work included (but were not restricted to) him being regarded as an "intrusive amateur [...] too closely related with a church [...whose] deployment of mathematics [...] was confusing" (Atkins 2003, p. 47). Similarly, the ideas of the psychologist Egon Brunswik were discarded, because "[c]orrelation and regression statistics had become Brunswik's indispensible tools of the experimenters' rival community, "correlational" or "differential" psychologists [...]. Lack of acceptance went hand in hand with lack of understanding" (Gigerenzer 1994, p. 57).

Similarly, reasons to do with pride or other emotions may hinder scientists from accepting novel findings. The "confirmation bias" or "myside bias" (cf. Nickerson 1998, Baron 2000, p. 195) observed in Psychology shows that people tend to cling to their beliefs even when presented with contradictory information. Also, when seeking information, there is a tendency to look for information corroborating rather than contradicting one's own beliefs. Such behaviour can lead to the rejection of novel findings (cf. Nickerson 1998). A paradigm shift, as explained in Section 1.1, may be deferred as a result.

In turn, singling out new findings enables the scientific community to better react to progress. With heightened awareness of new findings, the chances increase that their impact will be objectively recognized. The approach developed in this thesis cannot change the procedures in the scientific publication system or the decisions individual scientists make. However, it can facilitate the dissemination of new findings that otherwise might drown in the vast sea of scientific publications.

This enables a better if not optimal utilization of available means and resources. Of course, machines or scientists are not idle until a new topic is discovered. Yet, awareness of the new topic enables a better allocation of the resources. In particular, the new findings can render specific tasks, experiments etc. obsolete. Furthermore, scientific funding can be better organized.

Figure 4: The process of filtering documents in emerging topics from those in old topics.



Source: Own illustration
Notes: LDA: Latent Dirichlet Allocation, as explained in Chapter 5.

Based on this, the goal of this thesis is to develop a system that detects documents in emerging topics. A schematic view of the proposed system is given in Figure 4. Two approaches, denoted here as "LDA" (Latent Dirichlet Allocation, details given in Chapter 5) and "Rules", are combined to process a set of documents, the "initial dataset". In the end, the system presents a set of documents to the interested user (cf. the "result set" in Figure 4). These documents are called "emerging topic candidates": They deviate from the "usual pattern" but their actual status, i.e. the novelty, has not been verified yet. Such a deviation can be based on the publication behaviour or the missing similarity with previous topics. Separating them from the other documents reduces the document set that a human expert would have to inspect. The main goal is to significantly reduce its size while ensuring sufficient coverage of emerging topics.

Two approaches are implemented and tested:

- The LDA approach builds clusters and calculates their similarity based on textual features
- The rule-based approach detects single documents that deviate from the publication norm

The two complementary approaches are compared and combined in this thesis (cf. Figure 4). Additionally, an approach that links topics based on their textual and reference features is implemented and tested. Both features, text and references, can be directly influenced by the authors of a paper and both can be used to form connections between topically-related documents: The words and vocabulary an author chooses as well as the references he cites indicate the topical background of a document. Both parts can be interpreted as the results of a convergence process wherein the authors – independently of their whereabouts – gradually reach an agreement about the terms and references to be used in a specific topic. In accordance with this convergence process, MacRoberts and MacRoberts (1996) describe a journal paper as "the last in a series of often dozens of laborious and painful and continuously changing drafts in which authors and co-authors construct, reconstruct, and negotiate knowledge and in which outsiders, notably referees, colleagues, and editors, add their two-bits, often including references (usually their own) the author has never seen." (MacRoberts and MacRoberts 1996, p. 441). Based on their connection to the published document, these features will be called "internal features" in the remainder of this thesis.

In contrast to these features, there are other factors which are not fully under the authors' control. For instance, the journal chosen for submission may reject the publication. Thus, the authors can only partially determine the output source. Also, even though their decision for a specific journal can be based on its characteristics (e.g. the Journal Impact Factor, for a detailed definition see Section 3.2.3), they cannot manipulate them. Thus, there are certain external factors that influence the publication process, which may differ for emerging and established topics. In particular, certain publication behaviour might be "forced upon" the documents in emerging topics, e.g. due to various impeding factors affecting emerging topics that will be discussed in this thesis (see in particular Chapters 6, 8 and 9). The resulting discrepancies in publication behaviour, however, can be also used to identify publications in emerging topics. These telltale indications of emerging topics struggling to gain ground in the scientific landscape are the basis for the set of rules. The respective features are thus independent of direct "manipulation" or influence by authors, and therefore called "external features" in this thesis.

There have been many studies introducing automatic or more often semi-automatic approaches to detecting new topics in science. Most of them rely on a comparison with other topics in the existing scientific landscape and citation-based indicators (see, for instance, Price 1965, Small and Upham 2009, Kajikawa and Takeda 2009, Shibata et al. 2009a, Shibata et al. 2009b). In addition to a high manual or computational effort, the former demands a knowledge base that includes all the necessary information for many years. Only then is it possible to monitor the fine granular changes in topics and thus detect emerging ones. In particular, because scientific fields shift and change, the topics have to be compared across field boundaries. Thus, tremendous efforts are necessary to establish sufficient coverage. In particular, the huge amount of manual effort required, i.e. labour and time, is very costly. Furthermore, the results of manual approaches are not necessarily reproducible.

Morris and van der Veer Martens (2008) apply the metaphor of the blind men and the elephant to the analysis of emerging topics: each approach uses and consequently 'sees' only one part of the available characteristics of publications. Specifically, many approaches so far focus on citation analysis in any form (direct citations, bibliographic coupling, co-citation, cf. Section 3.3.3). Citations are – even unintentionally – biased for many reasons (see Section 3.2.1 and Chapter 8, MacRoberts and MacRoberts 1996, Bornmann and Daniel 2008). In particular, a publication's accessibility and visibility can influence its citation rate tremendously. Citation-based detection relies on the judgment and capability of the scientific community to identify emerging topics among the vast number of monthly or yearly publications. In addition to detecting a new topic, the scientists also have to acknowledge and use it to provide grounds for a citation. These requirements make high demands on the capability and flexibility of the scientific community.

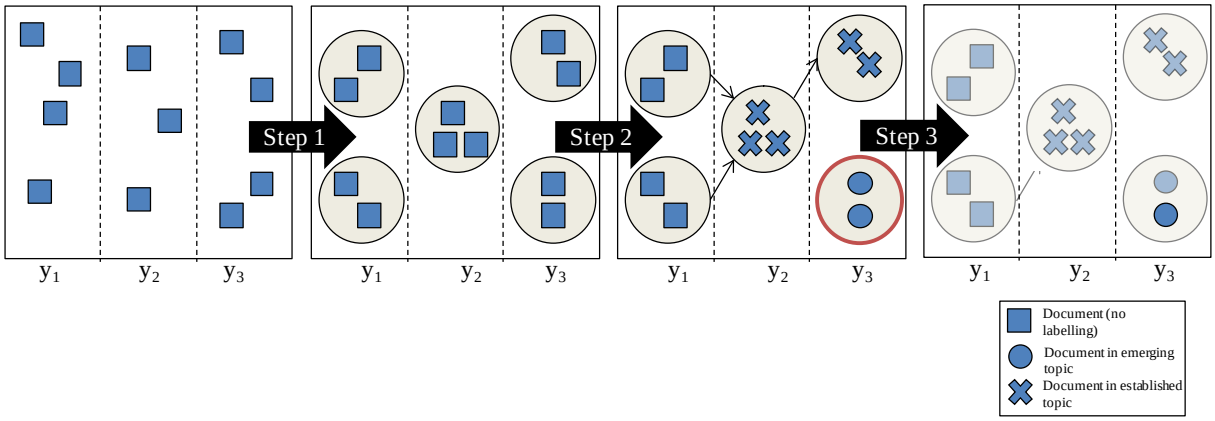
Moreover, citation analysis requires a time window. Citations need at least a couple of years to be made (Rinia et al. 2001, Glänzel and Schoepflin 1999) and a citation analysis can only be conducted two years after the initial publication date at the earliest. Thus, detection of emerging topics via citation analysis would have a time lag of several years (cf. Kajikawa and Takeda 2009) – a requirement that is counterproductive to the goal itself since, by then, the topics would no longer be emerging. Also, the fact that someone else has already cited the respective documents shows that the topic has been "discovered". In fact, as citation analysis relies wholly on the discoveries of others, the detection becomes obsolete.

In contrast, this thesis analyzes the characteristics of the publications themselves (and thus only information available at the time of publication) to identify indicators of innovative papers. The goal is to find indicators that could be used to separate innovative papers from the rest.

As a starting point, features are selected which are assumed to be suitable for this purpose. These features rely on characteristics of the publications and the respective references, journals or authors. It is important that the features can be calculated for all the documents involved and that their different values are comparable in any form. For textual features, this is achieved using a similarity function between the term distributions. More details are given later in Sections 4.3 and 5.2 and Chapter 9. Since interdisciplinarity is often named as one of the drivers of innovation, this feature is analyzed separately in more detail (cf. Chapter 7). Hypotheses for all the features are formed in Chapter 6 that describe how the feature values should differ for documents in new topics ("new documents") compared to those in established ones ("old documents"). For instance, one of the hypotheses states that new documents are published more often in multidisciplinary journals than in specialized, monodisciplinary ones.

The assumed setting for this thesis is a set of documents, *M*, from which documents representing new topics in science should be selected (cf. the "initial dataset" in Figure 4). These *M* documents[12] can cover several consecutive years, but at least two years are needed for the approach to work. In the example illustrated in Figure 5, the time span of three years is represented by the time periods y_1-y_3. The innovative papers are to be found in the most recent year, i.e. – in the example – y_3, which is also denoted as the observation year. In a real world application, this would correspond to the current year.

Figure 5: Topic clustering and subsequent connection of topics.



Source: Own illustration

For the observation year as well as the preceding years, the documents are grouped according to their topical relatedness (Step 1 in Figure 5). The topic models produced in this way are thus generated separately for each time period. For instance, in Figure 5, there are no overlaps in the topics for y_1, y_2 and y_3. Therefore, each year is treated separately to deduce the topics. Then, the topics of consecutive time periods are compared. Similar topics are "linked" to identify chains of topics and their evolution

12 The Appendix features a full list of all variable and parameter names used throughout this thesis. Please note the special usage of variable names for sets of instances and their size as shown on p. 206.

(see Step 2 in Figure 5). However, more importantly, those topics that have no predecessor can be identified (marked red in the graph). The respective documents are labelled candidates for new topics. A closer look at the exact approach will be taken in Section 4.3. Nevertheless, the main objective of this thesis is to identify parameters that could be used to

- group the documents by topic (Step 1),
- calculate the degree of similarity between topics in different time periods (Step 2), and
- make the result set more precise by applying the specific indicators of emerging topics (Step 3, rejected documents are greyed out)

In particular, the time span inspected needs to be evaluated: Are the predecessors of a topic only sought in the preceding year or is a longer time span considered? Of course, the chance of finding a predecessor increases with more topics and not all topics appear in two consecutive years. Yet the computation time and the error rate also increase if the time span is extended to include more years. This has to be considered when defining the time span for the approach (see Section 9.3.2).

Furthermore, the set of indicators used in the overall approach has to be determined. As mentioned above, different features and hypotheses are tested in this thesis. In the end, those that are the most useful for the detection of emerging topics are selected. In this way, specific characteristics can be derived for new topics. These characteristics can help to deepen the understanding of how new topics are formed and evolve.

1.3 Readers' Guide

The remainder of this thesis is structured as follows: There are three main parts which comprise an overview (the current part, "I. Overview"), the theoretical background ("II. Fundamentals") and the contributions of this thesis ("III. Contributions").

The first part finishes with Chapter 2, which gives a definition of emerging scientific topics and explores potential indicators. In particular, Section 2.2 provides the theoretical background for the explicit exclusion of citation analysis as an indicator of emerging topics.

The second part, "Fundamentals", explores the theoretical background of this thesis. Specifically, theoretical basics are explained in regard to bibliometrics in Chapter 3 and in regard to Machine Learning and in particular LDA in Chapters 4 and 5.

The third part presents the results of this thesis. A first, explorative attempt to assess possible indicators of emerging topics on the basis of a regression is conducted in Chapter 6. The role of interdisciplinarity for the development and discovery of emerging topics is explored in Chapter 7. A study that supports the theoretical background in Section 2.2 is presented in Chapter 8. In Chapter 9, the overall approach is assembled and calibrated in regard to the parameters. The chapter ends with an evaluation of the implemented system in a test environment. Chapter 10 examines the terms used in emerging topics. Finally, Chapter 11 summarizes the findings and gives an outlook to future applications.

Table 1 shows the author's previous publications and their role in this thesis. The thesis is based to varying degrees on these publications. In particular, earlier versions of Chapters 6, 7 and 8 have already been submitted, published or prepared for submission. The reader is referred to the respective publications for a more general setting of the analysis and results.

Table 1: Relation between this thesis and the author's previous publications.

Referred publication	Status	Usage in this thesis	Relation
Michels and Rettinger (2014)	Published (discussion paper)	Describes a previous version of the approach presented in Chapter 5	Preliminary studies for this thesis.
Michels and Schmoch (2012)	Published (refereed journal)	Cited on pp. 1 and 41	Serves as motivation because the observed increase in data due to better coverage makes the described system necessary.
Michels and Fu (2014)	Published (refereed journal)	Cited on pp. 27 and 40	Shows the influence of document types on citation analysis.
Schubert and Michels (2013)	Published (refereed journal)	Cited on pp. 24 and 43	Describes the influence of the publication source on the citation count of a publication.
Michels and Schmoch (2014)	Published (refereed journal)	Cited on p. 42	Shows the interrelation between monitoring publication behaviour and adaptation thereof by the observed authors.
Michels and Neuhäusler (draft)	Prepared for submission (refereed journal)	Resembles Chapter 6 in large parts	Shows the differences in bibliometric characteristics for emerging and established topics.
Michels (2013)	Published (conference proceedings)	Resembles Chapter 7 in large parts	Describes various methods to measure interdisciplinarity and its importance for emerging topics.
Michels (under review)	Under review (refereed journal)	Resembles Section 2.2 and Chapter 8 in large parts	Analyzes the interrelation between a topic's status and the citation count and timing of the associated documents.

Source: Own illustration

2 Emerging Topics and Their Indicators

The outline and implementation of a system to specifically detect emerging topics depends heavily on how these topics are defined and the inferred assumptions. However, even a common understanding of the term "topic" has proven difficult, in particular because the terminology changes with each new definition. In the following, a brief description is given of different levels of aggregation in science, i.e. topics, research areas and disciplines, where a discipline is the highest aggregation level for sets of publications. This overview is intended to foster an understanding of the approach and the granularity of its application.

Topic

According to the Oxford dictionary of English (2010, p. 1521), a topic is "a subject of a text, speech, conversation, etc.". In the following, a topic is defined as a common subject of at least two independent documents, i.e. – in this thesis – scientific publications. Thus, a minimum of two groups of scientists have to work on the topic independently of each other. This explicitly excludes algorithms in Computer Science that are only used by single research groups. Nevertheless, it allows high granularity in terms of topic size: Small topics are made visible but their birth and death rate is relatively high. Every topic introduced by one research group or author that does not arouse the interest of at least another person seems to be merely personal interest (or a "hobby horse") and not scientifically worthwhile. This holds in particular for the assumption made on scientific progress in Chapter 1. The involvement of different authors or research groups is thus necessary.

Specific to scientific publications is that the continuation of a topic is perceivable via the reference lists, even if the authors of a second publication do not use the same words or concepts. Therefore, even though the reference list might not be a clear indicator for the innovativeness of a publication, it can indicate the document's topic. Analogously to the textual part, the reference list is thus a mixture of topics (cf. Section 5.1, Leydesdorff 1998).

This thesis relies on the definition of "scientific specialties" by van den Besselaar and Leydesdorff (1996) in accordance with Kuhn (1973, pp. 10f) to define topics. In this case, a scientific specialty, or *topic* as it will be called in the remainder of this thesis, is an aggregation of an author network ("people working on the same set of research questions"), a methodological network ("using the same methods") or a reference-based, i.e. co-citation-based network ("referring to the same scientific literature", van den Besselaar and Leydesdorff 1996, p. 416). In agreement with McCain (1990), they state that the "communication within such a network is more intensive than the communication with researchers in other specialties" (van den Besselaar and Leydesdorff 1996, p. 416). In order to measure such communication or rather to use this communication to identify the respective networks, they use citations between journals.[13] Their study on Artificial Intelligence shows in addition the difficulty of a) deciding whether a scientific topic is still in an emerging phase and b) defining the scope of a topic.

---

13 For a discussion of establishing disciplinary boundaries via journal-to-journal citations see Section 3.3.4.

Research Area

A research area lies between a topic and a discipline in terms of its aggregation level and thus (typically) size. Similar to a discipline and in contrast to a topic, it is definitely persistent. Yet new areas might be spawned. Areas are collections of established topics that are publicly known and approved. Due to their long persistence, they have clearly assigned labels. Examples for areas on different aggregation levels are the Web of Science subject categories (cf. Section 2.4) and the 22 fields defined for the Essential Science Indicators[SM]14. The research areas of the latter range from Neuroscience & Behaviour and Agricultural Sciences to Space Sciences.

Discipline

Again using the definition given by the Oxford dictionary of English (2010, p. 408), a discipline is "a branch of knowledge, especially one studied in higher education". Thus, the hierarchical relations between a topic, a research area and a discipline can be illustrated in terms of the course of one's studies; first, a student chooses a *discipline* that he wants to study. After some general courses as an introduction to the discipline, he has to decide which areas he wants to specialize in. In the end, the final thesis is one of many *topics* the student picks in this area.

Several classification systems exist for scientific disciplines, especially for the use of educational studies. The list of fields of studies by the Statistisches Bundesamt offers a classification in 10 disciplines (Fächergruppen), which are in turn separated into 60 fields of study (Studienbereiche, Statistisches Bundesamt 2013). The specializations of these fields (270 in turn in total) are all on similar aggregation levels, but still differ notably in size. For instance, the discipline "Sports" has only one field of studies ("Sports, Sport Science") which covers two specializations ("Sports Education/Psychology" and "Sport Science"). Conversely, the biggest discipline ("Linguistics and Cultural Studies") has 17 fields of studies and 87 specializations, ranging from "Ethics" to "Computer Linguistics" to "School Pedagogics".

The "Classification of Instructional Programs" (CIP) by the U.S. Department of Education's National Center for Education Statistics (NCES) distinguishes between 47 disciplines ranging from "Visual and Performing Arts" to "Engineering".[15] The Joint Academic Coding System (JACS) is provided by the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency (HESA). It currently contains 88 disciplines[16]. The JACS classification is regularly reviewed and updated in order to "understand any new developments in the identified areas that may not have been reflected" in previous versions.[17] In the last update (to JACS 3.0), 12 areas had to be reviewed. Among others, Bioengineering was moved to a higher hierarchical level in order "to reflect that Bio-

14 http://sciencewatch.com/about/met/journallist/, last accessed on 2012/08/16.

15 http://nces.ed.gov/ipeds/cipcode/browse.aspx?y=55, last accessed on 2012/08/16.

16 The term "discipline" is used somewhat loosely in this context, but all labels at the second level in the hierarchy basically correspond to the definition of a discipline in this thesis.

17 http://www.hesa.ac.uk/content/view/1776/649/, last accessed on 2012/08/16.

engineering [...] is now a well-defined discipline in its own right".[18] In a similar way, Computer Sciences was split from Mathematics.[19] These examples show that, despite the above made statements, there can be changes in the order of disciplines or areas, but they are not as frequent as changes in the topic landscape. Disciplines have longer durability and cover several topics. Changes in topics might be reflected in the overlying disciplines. In this case, the changes have a permanent effect the overall structure. For instance a topic could be elevated to the hierarchical level of a research area/discipline. In other cases, a novel, overlapping discipline can better account for the frequent interchange between topics of previously disparate disciplines.

Based on the definition of a topic used above, the remainder of this section gives the definitions for emerging topics and features that are used in the course of this thesis. Possible indicators are discussed. In particular, Section 2.2 explains why citation analysis, a widely-used tool in bibliometrics, cannot be applied to detect emerging topics. The theories that are presented here are confirmed later by comparing the number and timing of citations of emerging and established topics (Chapter 8). The Section 2.4 depicts the data used in this thesis, while the final section relates the main points of this chapter to the remainder of the thesis.

2.1 Definition of Emerging Topics

As stated above, a topic denotes the conceptualisation of combined efforts, means and intellectual basis (references). The topic itself can be reified by the resulting scientific output, i.e. a set of scientific publications. Scientific publications and the accompanying bibliometric data are more informative than a summarizing text for the topic (even though this might be the more condensed form with no redundant information). Thus, the representation of a topic (immaterial) via the connection to scientific publications (materialized, even if in digital form only) is superior to a textual abstract. Based on this information, connections to other publications or topics can be traced, e.g. by the (re)use of vocabulary/terminology or references to former work. In this way, a network of topics can be constructed that spans several years and shows relations between topics.

More importantly, however, topics without connections or with only weak ties to earlier topics can be identified. It seems reasonable that topics without a predecessor are by definition emerging. This is the case in particular if a leap is taken in the development of an idea, such as a new treatment for a disease or the improvement or new application of an algorithm in Computer Science.

The basis for such a new topic is a discovery or the merging of previously unrelated topics. Basically, both forms of evolution result in findings that enable new technologies or applications capable of solving old problems. Topic development can be better illustrated by comparing it to the evolution of spe-

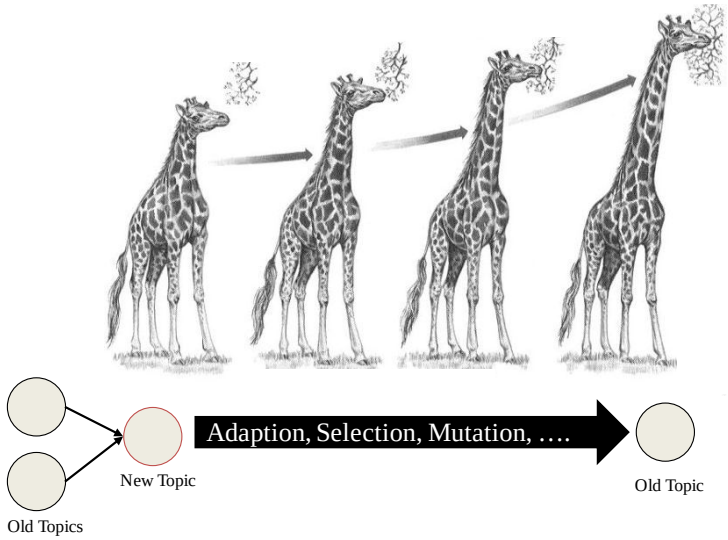
---

18 http://www.hesa.ac.uk/dox/datacoll/jacs3/Bioengineering.pdf, last accessed on 2012/08/16.

19 http://www.hesa.ac.uk/dox/datacoll/jacs3/Mathematics_and_Computing.pdf, last accessed on 2012/08/16.

cies (see Figure 6[20]). Every species evolves over time, in the same way that every topic develops with every new publication. The changes may be marginal in individual years, but a change is perceivable in retrospect, even for an outside observer. Branches of a topic, like the different races of a species, might adapt to different circumstances more or less successfully. However, a more rapid and drastic change in the set of topics and species can be achieved by combination or mutation. A mutation is a change in the genetics that introduces a new combination independent of hereditary. New gene combinations enter the gene pool, which allow new attributes or abilities. Thus, a former topic continues to exist, but on a higher level or in a different direction or manner which make the connection no longer important or even tangible. Swanson (1993) states that a scientific speciality "may fragment into new subspecialties or develop new relatedness with other older specialties – relatedness that may be unintended and unnoticed" (Swanson 1993, p. 619). In the same way, previously independent specialties might merge to form a new hybrid species. Similar to a new species in biology, a topic which results from the combination of other topics can also be assessed as "new". This merging does not necessarily occur between two equally developed or equally influential topics. Rather, Upham and Small (2010) write that a successful research front can be "absorbed" by other fronts if it does not grow in competition.

Figure 6: Development of a topic over time in comparison to real evolution.



Source: Own illustration, upper image downloaded from deskarati.com[21]

On a lower aggregation level, topics can emerge as one result of new ideas or findings. Johnson (2013) lists different sources of such ideas, most of which depend on lucky chances and/or a diverse stimulation. In the following, a closer look will be taken at four of these sources:

---

20 Actually, the illustration in Figure 6 depicts the theory of evolution according to Lamarck, which is an even better image for the development of scientific topics than Darwin's theory. Lamarck's theory represents a change based on intrinsic motivation (of a species or – transferred to this thesis – a topic), while the "survival of the fittest" in Darwin's theory depends on external selection criteria.

21 http://deskarati.com/wp-content/uploads/2012/10/lamarks-giraffe.jpg, last accessed on 2014/02/17.

1. The “adjacent possible”: One natural limitation of ideas is that they need the appropriate context and means in order to become applicable and acceptable. Babbage’s “analytical engine”, for instance, could not be constructed during his lifetime (Johnson 2013, pp. 46f). Thus, even though the first programmable computer had been drafted, the material needed for its construction was not yet available. Analogously, the applications and means need to be available for ideas to be accepted (Finger 2000, p. 305). Atkins (2003, p. 47) argues that – besides many other factors – only a modern view could help to put Mendel’s results in context with inheritance. Thus, progress in such an area might be impeded until other aspects “catch up”. In contrast, other inventions or findings seem to force themselves upon humankind, as they “appear” in different contexts independently (Ogburn and Thomas 1922). For example, the automobile was invented almost simultaneously by both Gottlieb Daimler in Stuttgart and Karl Friedrich Benz in Mannheim without prior consultation between the two. Kuhn (1973, p. 65) also refers to the frequently made observation of similar and concurrent discoveries and attributes it to the constant demand for change as soon as a paradigm is introduced.[22] A similar notion is the establishment of a general framework with gaps to be filled later. A good example of how the “adjacent possible” might be conceivable but still not tangible is the periodic table by Mendeleyev, which, at the time of its publication, contained gaps for elements yet to be discovered (cf. Boden 1994, pp. 81f).
2. Serendipity: Lucky chances have helped to unearth new facts but also permit new perspectives of established frameworks (cf. Grinnell 1987, pp. 31ff). Penicillin is one of the more famous examples of a lucky stroke as its discovery followed Fleming’s observations of a mould. Serendipity is the exact opposite of goal-oriented work. Also, by definition, it is difficult to trigger. Johnson (2013, pp. 111ff) names information exchange, distractions, browsing etc. as possible triggers of serendipity. In analogy to a poem by Roth (“Das Hilfsbuch”), a researcher might have begun work with a certain outcome in mind, thus acting in a goal-oriented way, but then dismissed that notion for another more promising route.
3. Falsity: Disruptive factors have been shown to enhance the creativity and outcome of research groups (Nemeth 1995, Cheruvelil et al. 2014). Also the results of erroneous experiments etc. are not necessarily setbacks (cf. Section 1.1). Considering the running example followed in this thesis, Mendel’s research was initially assessed by himself and his fellow scientists as a failed attempt to deliver a background for hybridization: The original experiments performed by Mendel were not meant to be an explanation of heritage, but a scientific foundation for hybridization. Considering only this aspect, they were a failure – and were also presented as such. Regardless of this, however, the results held the potential for the discovery of the rules of heritage. That they were the outcome of an operation with a completely different purpose is irrelevant. Quite the contrary in fact: Since errors in concepts and executions force researchers to think in new ways and adapt to new circumstances, they build one foundation for the paradigm shifts mentioned by Kuhn (1973, pp. 71, 77).
4. Exaptation: Established methods, means or tools from other fields can be adapted or misused to build something new. In nature, mutation, falsity and serendipity lead to evolution or the “next step” (Johnson 2013, p. 172). Exaptation is the natural consequence after accomplishing a new tool (via a mutation or serendipity) for new purposes. One example is Gutenberg’s usage of wine presses for book printing (Johnson 2013, pp. 168f). Exaptation can also be a “spillover” between disciplines or research groups, like in the case of Watson and Crick, who acquired ini-

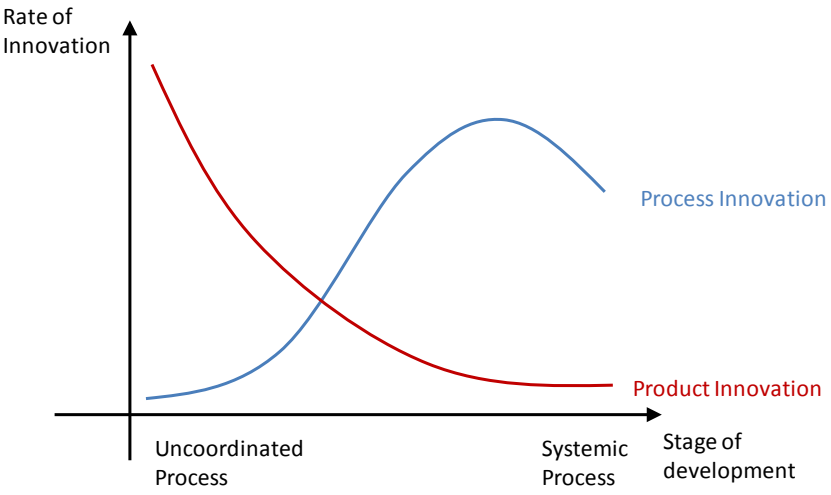
22 “[T]raditional pursuit prepares the way for its own change” (Kuhn 1973, p. 65).

tial input about the molecular structure of DNA from another research team (Atkins 2003, p. 80, Johnson 2013, p. 184). A study by Ruef (2002) also shows that organizations with a bigger and more multifaceted social network are more creative. The respective people collect ideas, findings, means etc. from various sources and contexts and are able to combine them. According to Boden (1994, pp. 75f), creativity is characterized by the ability to mingle existing ideas in a new way.

Exaptation is not the only source of novel ideas, but it integrates the other methods. Also, it enables the "gatherer" to outsource the (knowledge) production process to many heads. This allows him to focus on the combination of the means and tools at hand.

Analogously – and corresponding more to the typical image of a researcher – gatherers in science do not necessarily need a big social network. They are able to collect ideas and insights from various sources using publication data. There are – in theory – no restrictions in respect to the quantity or disciplinary borders for this process. However, exaptation is especially fostered when disciplinary borders are crossed. This is fully in accordance with the given definition of novelty that something new can be created by the adaptation of outside means and concepts.

Figure 7: Innovation and stage of development.



Source: Own illustration, adapted from Utterback and Abernathy (1975, p. 645)

When the topic starts, no "planning" is possible as the exact outcome, potentially applicable and derivable methods have not yet been determined (see left hand side in Figure 7). In this regard, the topics resemble products for which, at the beginning, "market needs are ill-defined and can be stated only with broad uncertainty" (Abernathy and Utterback 1978, p. 45). However, social needs or interests in combination with new findings leads to new topics (cf. Upham and Small 2010). As stated above, the outcome might be a topic that can develop independently and even evolve into a scientific discipline (Upham and Small 2010).

The progress in the beginning however is largely determined by the social setting of the topic. Investments have to be made by other researchers before entering or even understanding a topic, making the communication of findings, concepts and possibilities difficult (Callon 2000). Also, the knowledge production process and outcomes are instable in the beginning – a fact that can also influence the vo-

cabulary used by the researchers. More tangible might be the rivalry and the uncertainty in the network configuration, as methods, expectations and individuals enter and exit repeatedly (cf. Table 2). As explained above, rivalry or even "power struggles" might not only hinder the knowledge transfer from the inside of a topic but also from the outside, as the establishment and acceptance of a new topic is a highly socialized process. The process heavily depends on the way of dissemination, which in turn requires a strategic approach to ascertain stability and then growth (Weingart 2003, p. 47).

Table 2: Comparison of network configurations for emerging and stable topics.

	Emergent Configurations	**Stable Configurations**
Knowledge	Statements + instruments + embodied skills	Statements are information because embodied competences are duplicated
	Non-substitutability between codified knowledge and embodied knowledge	Codified knowledge and embodied knowledge are relatively substitutable
	Private knowledge: rivalry and excludability	Knowledge is public – i.e. non rival, non-exclusive – within the networks where it circulates
	Knowledge replication = laboratory replication	Knowledge replication = coding and replication of strings of symbols
	Local knowledge is generalized through successive and costly translations	The degree of universality of knowledge is measured by the length of networks
States of the world	List and identity of social and natural entities constantly reconfiguring	List and identity of social and natural entities are known
	States of the world revealed, ex post, through trials and interactions	All states of the world are known ex ante and the probability of their occurrence can be calculated
	Uncertain and vague knowledge uses	Uses of knowledge are predictable
Modalities of action	Programs only exist ex post, as the outcome of action and learning	Research programs (problems + operations) are defined ex ante and provide a framework for action (coordination)
		Rational expectations
	Cooperation is an obligatory passage point for action i.e. for translating identities and interests and for negotiating the content of knowledge	Cooperation is a strategy for cost and risk sharing or for consolidation of power position

Source: Adapted from Callon (2000, p. 203), translation by Callon

In the remainder of this thesis, emerging topics will always be differentiated from "established topics", i.e. topics, which have already reached the level of maturity described above. For the sake of simplicity, they might also be referred to as "new" and "old" topics, respectively. In the same way, to facilitate reference at the document level, documents in emerging topics will be called "new documents" or "new instances". The remaining documents will be referred to as "old documents" or "old instances".

"Innovation" in this context refers to the process of working in such an emerging topic as this – per definition – requires the application of new methods or ideas.

The topical relatedness of two groups of documents can be measured in various ways. In this thesis, one dataset is based on the Science Map report by the Japanese National Institute of Science and Technology Policy (NISTEP, Saka, Igami and Kuwahara 2010). In this report, Saka, Igami and Kuwahara calculated the overlap in research fronts with the preceding report. Other studies have used citations or references to map the connections between topics (cf. for example Small 2006, Shibata et al. 2009c, Hummon and Doreian 1989, Kajikawa and Takeda 2009, Verspagen 2007). However, citation analysis assumes there is already a sufficient degree of awareness of ongoing work on the topic and a time lag that makes citations between related works possible (cf. Section 1.2). Since the data in the NISTEP Science Map report was generated by overlaps with the previous report, this would – in a more general context – demand knowledge about previously established topics. In this thesis, the analysis of topics in former periods is one of many features on which the discovery of new topics is based. It is complemented by other features that rely on publication data. The following section explains in more detail why citation analysis cannot be used to discover emerging topics.

2.2 Citation Behaviour as an Unreliable Indicator of Innovation[23]

A scientific career is heavily influenced by the impact of the respective publications. An unread or unused scientific publication cannot foster the progress of science. Usually the impact of publications is measured via citation counts. Citation counts easily denote how often a publication is used – regardless of which way. Citations are counted regardless of their context or their quality. In this sense, scientists' attitude to fame is the same as celebrities' – bad publicity is better than no publicity.

Some studies even suggest that citations equal some kind of quality measure. But clearly, there has to be a differentiation made. For one, citations can be used to refute or criticize previous work. Also, general reviews might be cited more often even though or because they present a summary of earlier work rather than novel ideas (cf. MacRoberts and MacRoberts 1996). Other work is applicable in more than one context and is thus cited more often, e.g. Aksnes' paper (2003) on self-citations is frequently cited when the exclusion or inclusion of self-citations in bibliometric studies is discussed.

Various forms of citation-normalization have been introduced in bibliometric analyses to make up for differences in citation behaviour across scientific disciplines or fields, countries and even journals (c.f. among others Leydesdorff and Bornmann 2011, Leydesdorff and Opthof 2010, Leydesdorff, Zhou and Bornmann 2013, Zhou and Leydesdorff 2011, Zitt and Small 2008, Waltman et al. 2012, Glänzel et al. 2011, Zitt, Ramanana-Rahary and Bassecoulard 2005, Zitt, Ramanana-Rahary and Bassecoulard 2003). Sometimes citation rates are related to the average citation number in that field (field expected citations) or journal (journal expected citations). These values are supposed to relate the citation numbers to the "usual" value, which might differ by field or depending on the journal's popularity and market coverage. Normalization might succeed in cases where the bias can be corrected automatically

23 Parts of this section were submitted as part of a research paper (Michels under review).

by comparison with expected values, e.g. the average field citation rate. However, in the remaining cases, the distinguishing factor might not be identifiable or extractable. In particular, in the case of emerging topics, the normalization for the current "status" of a document's topic might prove difficult to implement.

The basis of an emerging topic in science can be a new discovery or a merger of former unrelated topics. In turn, this may be the result of serendipity or alternative creativity, thinking outside the box, or any other factors that can be designated a "lucky strike" (cf. Section 2.1). All kinds of evolution can lead to findings that enable new technologies or applications. The speed with which their impact matures in a field can vary (Dorta-González and Dorta-González 2013) and thus also the reception of the new topic in the field which is measured using citations.

Citation rates are usually treated as an indicator of a paper's quality. They are used as an external signal to judge the complex concept of quality. This corresponds to using audience ratings to measure TV programmes' quality, whereas, in most cases, high rates are the result of a "concept for the masses" combined with good placement (time and channel).[24] In line with this idea, a low citation rate could also indicate outstanding papers in so far as no one else is working on the same topic.

In accordance with the "concept for the masses" theory, Garfields's (1979) explanation for the many citations of a paper by Lowry et al. (1951) is supported by the statement that "nearly everyone" could use it. Therefore, citations are – at the point in time when they are made – merely a sign of ongoing research that can be easily adapted by other scientists. Arrow et al. (2011) stated that, in economics, citations are biased in favour of the largest subfields. On a similar note, citation counts are, according to Garfield (1979), "…nothing more, nor less, than a reflection of that [scientific] community's work and interests" (Garfield 1979, p. 364). In accordance with that, he also defines the number of citations a paper receives not as a measure of "importance" or "impact" but of "utility" (Garfield 1979). The problem is that the converse argument is not necessarily true, as there is no implication about the (potential) usefulness of uncited/unnoticed publications.

In contrast to Garfield, van Dalen and Klamer (2005) state that the uncited majority in science is not – as sometimes purported – mere "waste", but a necessary side-effect of creativity: "Eliminate waste and you eliminate the possibility of a rare, outstanding piece of work" (van Dalen and Klamer 2005, p. 399). The same rule applies to innovation in academia ("the winner takes it all"). According to the authors, trial and failure are thus the typical fare in science.

Of course, lower citation rates can also stem from other factors, some of which might be mere side-effects of publishing in a new topic. Johnston, Piatti and Torgler (2013) show that the citation rate is lower for theoretical (in contrast to empirical) and single-country publications. The former are hard to evaluate on a large scale (Johnston, Piatti and Torgler only show it for publications in the American Economic Review), but the number of countries can be more easily evaluated. The question is, however, whether the latter is perhaps also a side-effect of some other (hidden) characteristic of the paper.

---

24 Note that, in science, the target audience is both supplier and consumer (cf. Franck 1999), while in the TV programme simile, the audience acts as a consumer only.

Documents might be underestimated or less visible because they are cited less than other publications. This might be enforced by the Matthew effect (Merton 1968) – those that already have much (i.e. many citations) gain even more, those that have less gain nothing.

As in the case of Mendel, certain publications might be "overlooked" by the scientific community and thus not cited for a longer period despite their (latent) impact (cf. van Raan 2004, Costas, Leeuwen and Raan 2011). The innovative papers by Nobel Prize winners Hans Krebs and Barbara McClintock were rejected by the journal Nature (Kilwein 1999, Benos et al. 2007). However, Glänzel and Garfield (2004) showed that belated recognition in terms of citations is relatively rare (0.013%) and affects the scientific disciplines differently, concluding that the Mendel syndrome does not occurs as often as one might think.

Franck (1999) stated that time constraints make citations more and more superficial and that "advertising, public relations, and marketing" are therefore necessary tasks on the authors' (i.e. the suppliers') side to make up for this deficit on the readers' (consumers') side. He mentions that this is especially the case with a new theory for which "the attention received […] often differs from what it deserves after a second look" (Franck 1999, p. 54).

As such an "advertising" effect, Schubert and Michels (2013) showed that the choice of journal can influence the citation rate of a publication. Similarly, van Dalen and Klamer (2005) pointed out that the reputation of a journal can be seen as a signal for the scientific community. However, it is also a question how much authors can influence where they publish novel ideas – "avoidance of unconventional ideas" is one of the biased reasons for rejections of papers noted by Benos et al. (2007). They conclude that "avoidance of avant garde and controversial topics by reviewers and editors could hamper the advance of science" (Benos et al. 2007, p. 147).

Barber (1961) gives a number of reasons why scientists reject new discoveries, explanations and methodologies in Science. In particular, he lists

- Substantive concepts: Concepts that persist because the new findings are not understood or acknowledged as the old ones seem to be more explanatory or intuitive.
- Methodological Conceptions: The resistance to findings that stem from applying different methods or perspectives, e.g. the usage of "rather difficult mathematical deductions" by Mendel in Genetics (Iltis 1932 cited by Barber 1961, p. 599).
- Professional Standing: "[…] the dignitaries who hold high honors for past accomplishment do not usually like to see the current of progress rush too rapidly out of their reach" (Zinsser 1940, p. 105 cited by Barber 1961, p. 601).
- Professional Specialization: e.g. "medical specialists have a long history of resisting scientific innovations from what they define as "the outside"" (Barber 1961, p. 601).

The issue might be more profound for premature discoveries, for which Stent and Hook (2002) name two criteria for identification: 1) No impact is achieved and 2) "its implications cannot be connected by a series of simple logical steps to contemporary canonical knowledge" (Stent and Hook 2002, p. 84). Thus, in contrast to the aforementioned arguments, the scientific community does not refuse to acknowledge new findings, but rather they cannot acknowledge them, because there is still a missing link to ongoing work (cf. Grinnell 1987, pp. 45f as referred to above). However, it is difficult to tell

the difference and in most cases also irrelevant. More often, the (fellow) scientists are blamed for their stubbornness – as for instance in the case of Planck:

"There are two lessons here for our comprehension of the scientific method. One is that revolutionary ideas gather strength from resistance to continuous attack. Unlike in some other fields of human endeavor [...], in science a crazy idea is subject to constant attack, especially – really especially – if it overthrows an established paradigm. The second lesson is that old men [...] are not the best evangelists of radical science [...]. Like new mores, new paradigms become accepted only as old generations die" (Atkins 2003, p. 205).

Glänzel and Schoepflin (1999) considered a citation window of 2 to 3 years appropriate in the major (non-Social Science) disciplines and concluded that "a reasonable part of cited (citing) documents is covered in such a short observation period, and the usual citation-based indicators can be considered appropriate to measure the impact of published research results in these topics" (Glänzel and Schoepflin 1999, p. 43). It is questionable whether this also holds for emerging topics, which might suffer from lower visibility and acceptance in the scientific community. However, on a similar note, Pollman (2000) argues that the process of forgetting scientific literature, and thus its decay, is independent for all publications regardless of the publication source, the publication year and especially the field. This holds – according to his study – for all publications aged 4 years and above. Similarly, there is a "natural tendency" for groups to use an up-to-date knowledge base and more current findings (Amat and Yegros Yegros 2009, p. 450).

In the course of this thesis, differences in citation rates and windows between emerging and established topics will be elaborated in order to highlight the necessity for other indicators to detect emerging topics (see Chapter 8). The remainder of this chapter examines possible indicators and the bibliometric databases used to apply the approach.

2.3 How to Find Them

Two approaches were used in this thesis to detect emerging topics:

1. Topic clustering to detect clusters that bear little resemblance to former topics and
2. Feature-based selection of "deviating" documents.

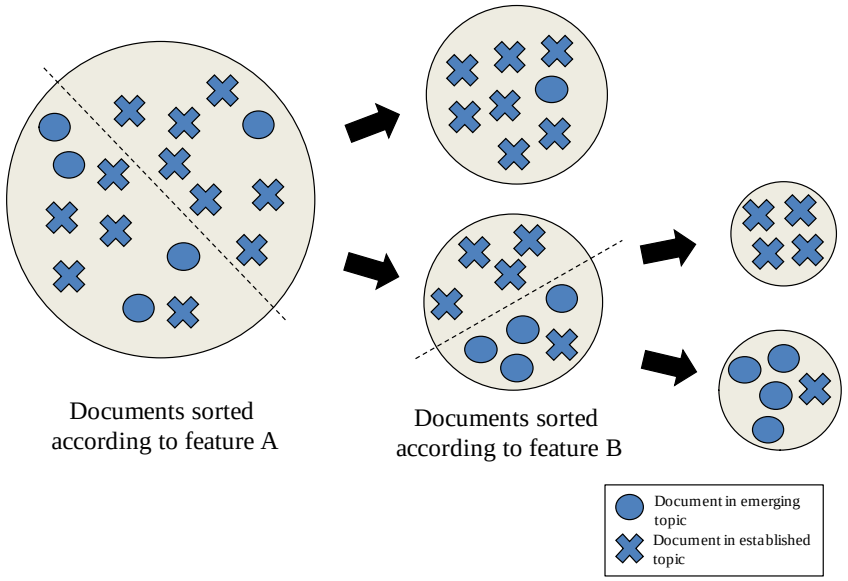
The relation of the two approaches was shown in Figure 4 in Section 1.2, where the selection of emerging topic candidates was explained. This section focuses on the rule-based approach that complements the cluster-based method.

So far, the term "features" denoted those characteristics of publications that are used to identify emerging topics. In general, the term encompasses all the characteristics of a publication that can be calculated in a standardized way; for example the publishing journal, the number of authors and the publication year. However, only some of these features can be used in the rules on which the system of this thesis is based. In particular, feature values of different instances must be comparable. Only then can they be used to separate specific document groups. Thus, features with a very high or low variance are ineffective. For example, neither a constructive division nor an understanding of the emergence process of topics can be gained by singling out the documents written by a particular group of authors.

Similarly, a distinction by publication year covers too many (if not all) documents in a dataset and gives no indications of the characteristics of emerging topics. A feature functioning as a differentiator must have a variety of feature values across the dataset.

In general, numeric features have many advantages. Instances can be compared and average values – and thus standard values for e.g. disciplines – can be calculated. Also, so-called threshold values can be introduced to separate a dataset according to values above or below a certain value. These threshold values are the foundation for the rule-based approach. Typically, rules can be derived with the help of the features that use the feature value and basic mathematical relations (equal, smaller/larger than, between) to separate sets of documents. This can be compared to skimming or cutting, where either the skimmed portion is eliminated (if it represents old documents in the majority) or selected for further processing.

Figure 8: Example of rule application for splitting a dataset.



Source: Own illustration

Figure 8 illustrates this process. First, the documents are sorted or arranged according to their feature values.[25] A threshold is introduced (represented by a dotted line) that separates all the documents with feature values above the threshold. As the threshold cuts off the majority of old documents, the process for detecting new documents continues with the other set. The respective documents can then be sorted and split based on another feature. The final result is the same whether the process is performed sequentially or in parallel.

The rules either target the new ("skimming") or the old documents ("cutting", where the residuum is used). Of course, each rule can be reversed to target the other type of document. For example, a rule selecting old documents with a feature value higher than 1 can be rewritten to skim new documents with a feature value of 1 or less. There will be many examples of this since the method determining

25 In the figure, this is represented in 2D and all the available space is utilized, even though, in a realistic scenario with simple numeric features and a coordinate system, the documents would be lined up.

the rules results in skimming as well as cutting rules, but the final approach is always applied to the document set containing more new documents and thus uses skimming rules only.

The following two subsections deal with the background of the two kinds of emerging topic detection used in this thesis: First, possible features for the rule-based approach are explained. After that, the background of topic monitoring is explained including the possible stages of a topic and their indicators. Being able to distinguish these development stages is the motivation for clustering and linking topics in this thesis.

2.3.1 Possible Indicators

Similar observations can be made for publication numbers as for citations: Publications in relatively unknown journals cannot be treated as equal to those in popular journals. Yet superficial studies lump them all together in one bin. Furthermore, sometimes the different preconditions in the scientific fields are ignored. For instance, conference proceedings publications have a different standing in Computer Science than in other fields; they are more important and acknowledged than in other fields and in some cases even outweigh journal publications (Michels and Fu 2014). But publication studies seldom acknowledge these differences.

Both metrics, citation and publication numbers, have thus to be used wisely when assessing the impact of a particular paper or topic. In particular, citation numbers should not be confused with quality. The goal of this thesis was to focus on alternative indicators.

Small and Upham (2009) performed a case study to determine the characteristics of an emerging research topic. Besides citation analysis, interviews with authors of highly cited papers in the topic give indicatons of the potential drivers of emerging topics – funding from industry and new interdisciplinary teams to name only two. Funding from industry provides a very clear indicator of the need for solutions in this topic (Stifterverband für die deutsche Wissenschaft 2013, Grupp 1998), which goes hand in hand with the definition by Kuhn (1973, pp. 71, 77). These findings are corroborated by a later study, in which 21 out of 26 highly cited papers in the emerging topic “Organic thin-film transistors” start by referring to possible applications (Upham and Small 2010). Kajikawa and Takeda (2009) measured the stage of an emerging topic by the journals that were used for publication and found differences in the publication strategies for applied and basic research – an interesting notion if access to an appropriate journal classification is available.

A further driver for the development of an emerging topic might be its applicability, not only in industry, but in various research technologies. Leydesdorff and Rafols (2011) explain this in their study on “siRNA” research. Guo, Weingart and Börner (2011) identified an increase in interdisciplinary research as another key factor. Other factors, according to their study, were the entering of new authors at the beginning and word bursts at the end. These word bursts are based on the frequency of unique words over time (cf. Mane and Börner 2004). Their approach does not take into account that vocabulary might be highly volatile at an early stage of a topic, yet it provides an overview of the spread of a topic in the scientific community.

The characteristics presented above suggest that documents in emerging topics differ in their publication behaviour. In particular, stronger ties to applied technologies can influence the choice of publication source and references. Larger author teams in established topics indicate that collaboration is hindered at the beginning of a topic. These factors thus formed the basis for the analyses (Chapters 6 and 9) to deduce indicators distinguishing emerging and established topics.

Small (2006) tracks the development of topics by building co-citation clusters in overlapping time periods using single-link clustering. First, the publications themselves are clustered. These clusters are again clustered repeatedly until the desired aggregation level is achieved. Clusters existing in different time periods, "threads of continuity", are called "cluster strings" (Small 2006). Small (2006) differentiates the forms of cluster development according to the distribution of the respective publications over time: They can grow or diminish, splice, merge, appear or disappear. Similar things are done in the trend analysis in this thesis. To apply Small's approach, the time period and the time splits first have to be determined. In this thesis, the time splits were set to years. The time window of the analysis was determined in the parameter estimation in Section 9.3.2.

Small notes that a cluster's origin may be merely a special issue in a journal and excludes them via a measure called "in-group citation". Parallels with this approach can be found for the detection of emerging topics in the NISTEP data, which is introduced in Section 6.3.1.

In this thesis, topics in an early stage are analyzed which might disappear again if they do not assert themselves. In contrast to that, Glänzel's (2012) definition of emerging research topics covers only topics which "have already reached a certain critical mass [... and] have strong links to their 'mother fields'" (Glänzel 2012, pp. 196f). In this thesis, the absence of links to earlier topics qualifies a topic as an emerging one. This is independent of the further development of the topics themselves. Other work has focused on manual efforts to determine the clusters' foci and connections (see e.g. Takeda and Kajikawa 2009).

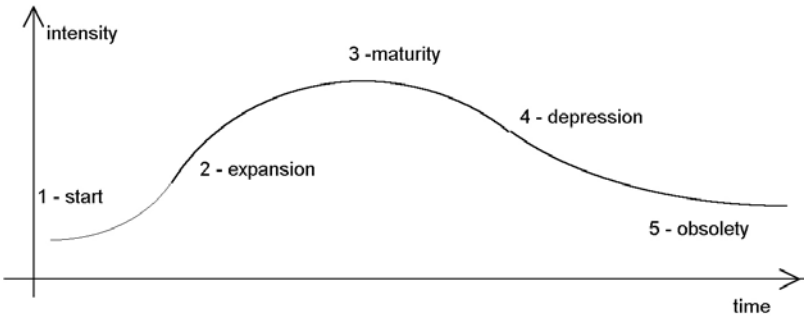
The identification of specific deviant characteristics of emerging topics was one way to implement a system that can detect emerging topics. Another possibility was to use the approach by Small (2006) mentioned above to create an overview of the topical developments of a document set first and then to extract those topics that come into existence during the analysis period. Such an approach was developed in this thesis, but it relies on inidcators different to those selected by Small (2006). For a better interpretation of the approach, the next section shows how topic development can be monitored and assessed in respect to the development stage.

2.3.2 Topic Evolution

After having created a set of topic clusters (for details on the implementation in this thesis see Section 5.2) the development stage of a topic can be assessed with the help of different features. The clustering in different time periods and the connection of these clusters also allows the illustration of the evolution of a topic for better assessment (cf. Small 2006).

The simplest form of an evolution monitoring is the calculation of publication numbers over time for a specific topic. In this case, growth or decline of a topic can be measured. Cahlik (2000) provides a classification in five different development stages of a topic according to its publication numbers (see Figure 9).

Figure 9: Stages of topic development.



Source: Cahlik (2000, p. 384)

It should be considered that the publication count is an easy to calculate metric if the topical boundaries are already established. Typically, pre-defined classification schemas enable the calculation of the publication number per topic. However, in a real-time[26] assessment, an established classification would not cover the emerging topics.[27] Thus, the assessment of such topic developments is only an option after the detection of the topic. Still, the discussion of the theoretic development stages can provide an image of the actual emergence and the ideal development process of a topic. After all, the "institutionalization" of new topics is a necessary precondition for their establishment (Thompson Klein 1996, with reference to Chubin 1976, p. 455).

Topic networks, in which the documents are connected in as well as across topic limits, allow the measurement of the density and centrality of the topics (cf. e.g. Coulter, Monarch and Konda 1998, Courtial 1994). The density represents the degree of connectivity in a topic, while the centrality measures the number of links with other topics. With the help of these metrics, the scientific topics can be divided into four groups according to the values for density and centrality (Callon, Courtial and Laville 1991, Cahlik 2000, Cobo et al. 2011, see Figure 10):

- Topics with both high density and centrality are very well developed and also used in other topics (upper right corner in Figure 10). Therefore, they are called "motor themes".
- Topics with a high centrality but low density lack the expertise of "motor themes" (lower right corner in Figure 10). Thus, they do not provide highly developed expertise for other topics but merely (basic) knowledge that can be easily transferred to other contexts.
- Low centrality combined with a high density hints at a high level of specialization that cannot be used in other topics (upper left corner in Figure 10). This might also correspond to former central topics that "have been progressively marginalized" (Callon, Courtial and Laville 1991, p. 166).

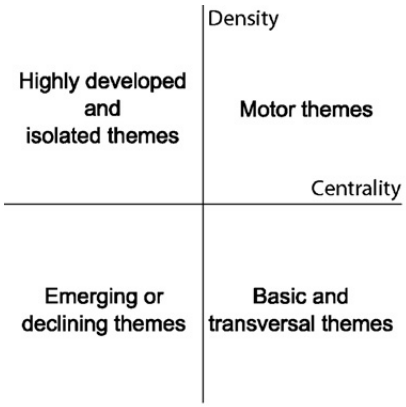
---

26 The term "real-time" is used in this thesis to denote any analysis without a (necessary or implied) time-lag.

27 And if it would cover them, it would render any method for their discovery obsolete.

- Peripheral and undeveloped topics (Callon, Courtial and Laville 1991), i.e. the ones with (yet) low density and centrality, might be categorized as "emerging or declining themes" (Cobo et al. 2011, p. 151), corresponding to topics in the initial or final stages of the topic evolution (lower left corner in Figure 10).

Figure 10: Categorization of topics in respect to their density and centrality.



Source: Cobo et al. (2011, p. 151), adapted from Cahlik (2000)

The system developed in this thesis targets this final group of topics, the emerging or declining ones. Topics with some kind of publication "history" that have diminishing values for both parameters (density and centrality) are per definition "disappearing themes". Noteworthy is, that in this case emergence and disappearance are equal in the measured metrics. This also reflects the fact that not all emerging topics develop a stability to evolve to another maturity level before disappearing again.

As was already mentioned in Section 2.3.1, numerous features can be used to create a network of documents or document clusters, but also determine the expressivity of the resulting map. The restriction to citations adds not only an undesirable time component but also other effects already mentioned in Section 2.2.

Cobo et al. (2011) use the *Inclusion Index* to look for overlaps in Keywords V in two themes k_1 and k_2 from different time periods (in their case publication years, see Formula (1)). By using this index, the establishment of connections is restricted to those themes that share keywords. Furthermore, they measure the stability of a theme that is covered in at least two consecutive periods with the Stability Index (Small 1977, Braam, Moed and van Raan 1991, Cobo et al. 2011, see Formula (2)). Basically, the Stability Index measures the number of keywords V that survive the transition of the theme from one time period to the following.

$$Inclusion\ Index = \frac{V_{k1} \cap V_{k2}}{\min (V_{k1}, V_{k2})} \quad (1)$$

$$Stability\ Index = \frac{V_{k1} \cap V_{k2}}{V_{k1} + V_{k2} - (V_{k1} \cap V_{k2})} \quad (2)$$

Cobo et al. (2011) investigate upon the thematic evolution of themes in the field "Fuzzy Sets and Systems". By applying the Inclusion Index to the themes created by a co-word analysis (for details see Section 3.3.6), they can illustrate the evolution of a theme. Edges between these themes show the development, while their thickness parallels the value of the Inclusion Index, so that topics that have a

higher number of common keywords are connected by a thicker edge. In accordance to the definition and assumption used in this thesis, if a theme has no connection to any theme in a previous period, this theme is an emerging one. However, the approach in this thesis uses additional features to establish the topic clusters and the respective connections.

In the work by Cobo et al. (2011), some topics that appeared before 2005[28] have no connection to a theme in a previous or following period. These topics emerged but were not accepted in the scientific community, namely:

- Relations[29]
- Fuzzy-Mapping[29]
- Entropy
- Necessity-Measure
- Intuitionistic-Fuzzy-Set

There are some limitations of the approach that could cause this: On the one hand, the topics are only built in a very restricted document set, namely those documents published in two major journals in the area of "Fuzzy Sets". Furthermore, the topics and connections are established solely based on shared co-words, but as mentioned above, the vocabulary as well as the categorization of the documents in an emerging topic is more fluctuant in its first instantiation as in later time periods (cf. Chapter 10). However, besides reasons caused by the approach itself, this could also be the observance of an actual "survival of the fittest" in research, i.e. the survival of only those topics that arouse interest and are thus adopted by other research groups. The necessity for the system developed in this thesis arises from the fact that the respective selection process does not only rely on objective reasons (cf. Chapter 1).

Mann, Mimno and McCallum (2006) use a variant of the Journal Impact Factor (see Section 3.2) to measure the impact of a specific topic by its citation number relative to its publication number. Other indicators have been transferred from journals to topics as well for this purpose. The accessibility of comparable indicators and the visualization of topics is surely a valid mean for understanding a topic or its development. Still, all things said about indicators and especially the restrictions concerning citation analysis apply and should therefore be taken into account (cf. Sections 2.2 and 3.2.1).

Rzeszutek, Androutsos and Kyan (2010) use a combination of LDA and Self-Organizing Maps to group and track topics in online documents over time. LDA is used in this thesis as well to cluster documents. However, the bibliometric data used in the LDA approach in this thesis contains additional information to that available for online documents.

As already mentioned above, the approach presented in this thesis is basically one form of outlier detection. In this case, documents that do not fit in the normal publication pattern or scheme are identified. For this purpose, a definition of the "normal" scheme of publication is important. As is the case

---

28 In total, five time periods were analyzed: 1978-1989, 1990-1994, 1995-1999, 2000-2004 and 2005-2009.

29 No connection to foregoing topics could be found as the respective topic was discovered in the first analyzed period.

with the approach presented here (cf. Chapters 6 and 9), the scheme or pattern might differ for scientific fields. Furthermore, the pattern must cover all related characteristics of the publications. In this thesis, it is derived from reference lists, vocabulary, author names and bibliometric indicators, so that a distinction for new publications is measurable. In particular, in the approach presented here, the distance in years to the reference lists and the compound of a reference list, i.e. the mixture of topics in a reference list (cf. Leydesdorff 1998), is compared. The underlying assumption is that novel topics might be based on more diverse topics, more fundamental research, older work or similar – at least in comparison with already established topics that can refer to ongoing/recent research and have a more focused reference list.

By trend the number of research topics increases with the career status of a researcher (Horlings and Gurney 2012), which is a natural deduction when thinking about the research stages ranging from PhD student to professor; with increasing career status, the possibilities to follow multiple topics at the same time increase. Thus, considering the author names in the approach can help to delimit the topic clusters or calculate the probability of an emerging topic. Besides the career status, there are other factors that might influence the chance that a researcher starts a new topic. Incentives could be his current career outlook, his former research, his current status and reputation or the development status of his current topic. The latter might be instantiated by a topic that does not yield enough problems, lacks perspective or attention or in other words, does not trigger enough citations for the involved researchers.

On a similar notion, a transition in topics might be observable if authors with different scientific backgrounds work on a topic, yet have not collaborated before. In this case, the approach in this thesis detects a lower similarity due to the low overlap in author sets with former topics. This might result in a missing connection to former topics and thus the labelling as emerging topic.

In the case of this thesis, a mixture of the presented approaches for topic monitoring is applied:

- first, the topics are connected over time to form strings of topics evolving over time for those that have at least one predecessor
- second, of the topics for which no suitable predecessor(s) could be found, the individual features like reference lists, authors etc. are analyzed for outlier detection.

Thus, by trend a high threshold for the similarity of topics in the first part is important. In this way no emerging topics are excluded from the remainder analysis just because they share a certain set of terms, authors or references with a former topic.

2.4 Data Used in this Thesis

Even though the approach presented in this thesis can be adjusted for any kind of documents in theory, the focus was restricted specifically to new topics in science. Thus, only scientific publications were used in the remainder of this thesis. In this way, the approach can benefit from specific features of scientific publications, which are explained in more detail in the following.

One advantage in bibliometric databases is the standardized form of (meta) data for the publications, in particular their title, authors, output source, type of document etc. In the beginning of bibliometrics, the data had to be aggregated manually from single journals. The introduction of ISI's Science Citation Index facilitated the use of these data as it was the first structured collection of them (Malin 1968, cf. Section 1.1). Therefore, these data are presented uniformly in the bibliometric databases nowadays. The standardized representation allows for an automatic analysis of the data – this as well as the consequential repeatability of the results enables the systematic analysis of publication data and is the foundation of bibliometrics.

The textual body of a scientific publication contains three components which represent in most cases the only features that are not meta data per se in the bibliometric databases: title, abstract and text. Titles for textual documents are common in general, but are obligatory for scientific publications. The usage of abstracts on the other hand varies for publication types. In the case of scientific publications, abstracts are usually detached from the textual body and provide an overview of the conducted research. In particular, the abstract should (in a good scientific publication) at least include the major findings or novel ideas encompassed in the document. In contrast to the main body, it does not contain references to former work.

Most documents have a list of keywords, but their selection depends on the publishers or respectively journals. Their specificity can vary, as the assignment is made either by the authors or the publishers according to a fixed list or freely.

In contrast to types of meta data for which there is only one correct value, some meta data can have different values in different bibliometric databases. For instance, the journal title of a publication has only one correct value. In contrast to that, the scientific discipline of a document can vary, as it is assigned by the database providers. In most cases, the journal in which the publication appeared determines its final classification. However, the information might vary for different databases. Such differences will be discussed later in more detail when the specific databases used in this thesis are presented (Sections 6.3.1, 7.2, 8.1, 9.2 and 10.2.1).

There are various publication types, for which in all databases articles and conference proceedings are certainly the most common and frequent ones (cf. Michels and Fu 2014). The focus of this thesis lies on these document types. Even though the document type assignment should be univocal it can differ for bibliometric databases (see e.g. Harzing 2013). The automatic inclusion of vast amounts of documents and the respective document type assignment seems to be error prone. Again, a closer look at error rates and discrepancies in the databases will be discussed later (see Section 3.1).

In this thesis, different bibliometric databases are used to train and test the approach. The main databases used were Elsevier's Scopus and Thomson Reuters' Web of Science (WoS). Both databases were implemented in an in-house Oracle SQL database. This allowed for more structural querying and aggregated results. The schema of both databases is similar in main parts (cf. Mallig 2010), i.e. there is only little information that is exclusive for one database. However, each database has specific qualities, advantages and disadvantages that were considered when a dataset was created. For instance,

proceedings for the main conferences in Computer Science appeared to be better covered in Scopus. Thus, Scopus was used to create a bibliometric dataset covering conference tracks (see Section 9.2.1).

As already mentioned before, the main parts of a bibliometric database are scientific documents for which at least the following meta data are given:

- Title
- Author(s)
- Source, e.g. journal
- Publication year
- Abstract
- Keywords
- Citations/References (in most cases linked with other documents in the database)

Note that bibliometric databases are mainly about the meta data of the documents. So far, access to the full text of a publication is either left out or handled separately (e.g. the online version of Scopus offers restricted access to full texts for registered users). The bibliometric databases are concerned with the overall performance and structure in science, not the in-depth information of individual documents.

Both databases assign unique IDs to the documents. These identifiers are also used for cross-references, for instance when one document in the database cites another. The publication sources as well receive an ID. Sources can be journals, conferences or books with focus on the former two types. The Book Citation Index was launched by Thomson Reuters in October 2011,[30] but is not included in the in-house implementation or this thesis. Elsevier started the "Scopus Books Enhancement Program" in the beginning of 2013, but the effects are also not (yet) perceivable.[31]

Scopus was launched in 2004 by Elsevier. It currently contains 41 million records in the online database.[32] In comparison, Thomson Reuters' WoS covers more than 12,000 journals as well as 148,000 proceedings.[33] At a first glance, when comparing these two databases, it seems as if by far WoS outperforms Scopus in terms of document numbers (Figure 11). Only the number of conference proceedings is higher in Scopus but this does not counterbalance the lower numbers for (articles and) other document types. However, in an analysis over the years covered (here publication year of the documents), Scopus has a broader coverage for more recent publication years than WoS (Figure 12). Thus, the choice of the database depends on the individual analysis.

---

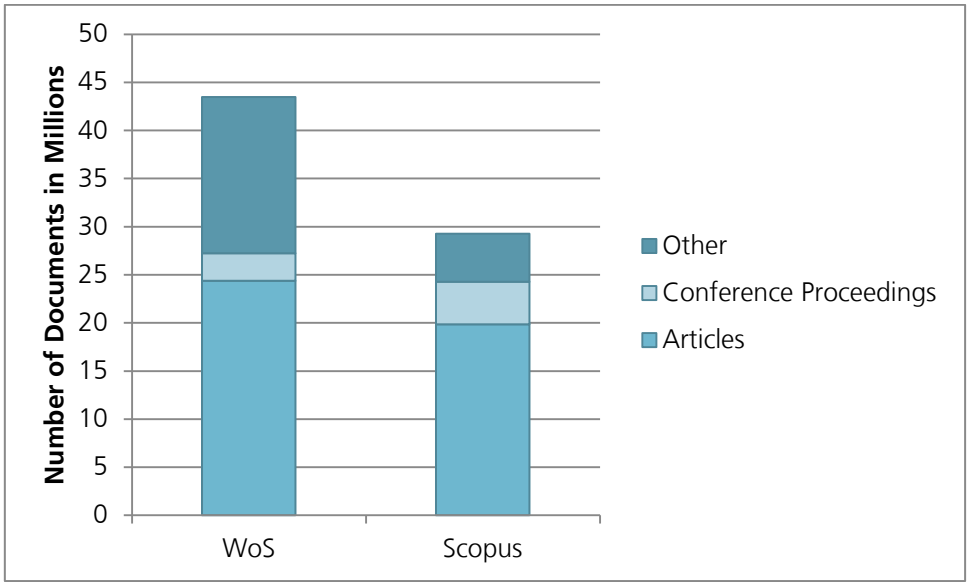
30 http://thomsonreuters.com/content/press_room/science/book-citation-index-launches, last accessed on 2013/08/20.

31 http://www.info.sciverse.com/scopus/scopus-in-detail/facts/, last accessed on 2013/08/20.

32 http://www.info.sciverse.com/UserFiles/2508.SciVerse.Scopus_Facts_Figures%28LR%29.pdf, last accessed on 2013/08/20.

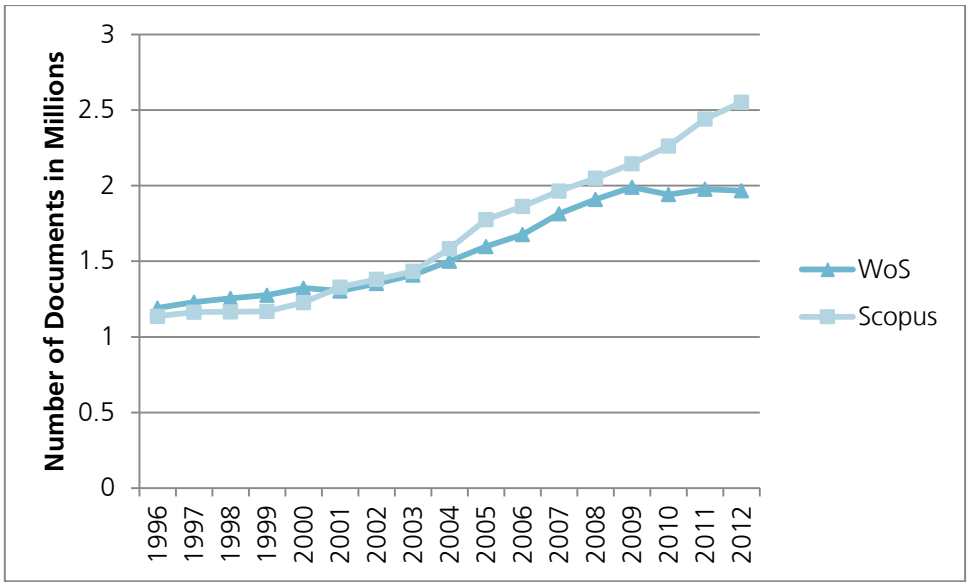
33 http://thomsonreuters.com/products/ip-science/04_064/web-if-science-factsheet-SSR0810070.pdf, last accessed on 2013/08/20.

Figure 11: Number of documents contained in the in-house database of WoS and Scopus.



Source: WoS, Scopus, own calculations and illustrations

Figure 12: Number of documents per year in the in-house database of WoS and Scopus.



Source: WoS, Scopus, own calculations and illustrations

The creation of an adequate testing dataset required knowledge of actually emerging topics and thus was most problematic. Even in retrospect, the creation of such a bibliometric dataset via the definition of search terms and other query strategies always resulted in disputable outcomes. Because of that, the created datasets used alternative methods to discover emerging topics, but were used as a benchmark or Gold Standard to evaluate the methods and processes developed in this thesis. In one instantiation, a dataset based on conference proceedings was used that included the track titles of the conferences for the respective publications (cf. Section 9.2.1). With these titles, the tracking of topics and in particular the identification of new topics was possible.

Another dataset covering a distinction between emerging and established topics was generated via citation analysis (cf. Section 6.3.1). In this way it could be tested if the results of citation analysis can be reproduced by the alternative procedure and thus detected independently and on time. However, as stated above, this was merely done to provide some kind of guide value for the tested approach – citations were still not used in the approach itself.

The dataset that was used in the main part of the evaluation (Section 9.5) was extensional information for the WoS database, where document IDs were labelled as belonging to a new topic or not.[34] This information was added to the in-house database to systematically evaluate the approach in the end. Its foundation was a document collection created by the NISTEP as a basis for a report on hot research topics between 2003 and 2008 in the WoS (Saka, Igami and Kuwahara 2010). The report data encompassed approximately 50,000 documents (articles and reviews), which were the 1% top cited papers in 22 fields. New topics were identified for the years 2007 and 2008. For more details about the creation of the research fronts and the identification of emerging research fronts, the reader is referred to the report by Saka, Igami and Kuwahara (2010). More details on these data are also given in Section 6.3.1.

Therefore, overall three databases are distinguished in this thesis: Scopus, WoS and the NISTEP data. However, as explained above, the NISTEP data are an annotated excerpt of the WoS data. Restrictions on the dataset, e.g. to specific years or categories, are explained in the section in which they are used (Sections 6.3.1 and 9.2). An additional random sample set of the WoS was used in the analysis of the term usage in the topics in Chapter 10.

2.5 Summary

This chapter provided an overview of different definitions of emerging topics and the associated indicators used for their identification. In particular, the advantages and disadvantages of applying bibliometric indicators were discussed. Caution is necessary as discrepancies in the databases can lead to varying results. Thus, a sophisticated database selection is mandatory.

Furthermore, this thesis discourages the usage of citation analysis to detect emerging topics for various reasons. The theoretical background underlying this approach was given in Section 2.2. This is elaborated in more detail in Chapter 8.

The remainder of this chapter focussed on monitoring topic development and the necessary features. The links between former studies and this thesis were shown, which will be elaborated in the following sections. Up to now, the components of the developed system were always presented in relation to each other or as a single system, so they are discussed separately in Part II. Hence, Chapter 3 deals with bibliometric analysis in general, so that the system's features can be placed in context. Chapters 4 and 5 enhance the understanding of Machine Learning approaches and the respective parts in the system developed in this thesis.

---

34 This is a simplified view of the data used in this thesis. The original dataset encompassed much more information and was not divided into new or old topics as such but instead into rapidly developing, new and other research fronts. All topics that were not labelled "new" were marked "old" for the purpose of this thesis.

II Fundamentals

3 Bibliometrics

Bibliometrics is the statistical analysis of publication data in science (cf. Havemann 2009, p. 7). According to its initial definition, this includes the mere descriptive or quantitative analysis of publication data (cf. Pritchard 1969). Indeed, publication counts and citation analysis constitute a huge part of bibliometrics (see e.g. Narin, Pinski and Hofer Gee 1976, Potter 1981, Broadus 1987).

The analysis can encompass single documents, authors, research groups, countries or any other defined set of the aforementioned (e.g. fields) and is in most cases of comparative kind, i.e. it is conducted for a number of subjects which are then compared in terms of the bibliometric indicators (see e.g. Wallin 2005, Bar-Ilan 2008, Thompson, Callen and Nahata 2009, Alonso et al. 2009, Garfield and Sher 1963). Typical bibliometric indicators are the absolute or proportional number of publications, citations or variations and combinations of the aforementioned in a specific time frame.

For instance, the influence or impact of a set of publications can be measured by the total number of citations they receive (for more details about citation metrics see Section 3.2.1). Micro-studies might be restricted to single specific publications, while macro-studies usually choose a set of publications from specific countries, years, subject categories and/or institutions.

Bibliometrics is enabled by the written instantiation of scientific communications, i.e. publications. Havemann (2009, pp. 7ff) distinguishes between the collective side of the scientific production process where the publications are shared freely,[35] and the competitive side where each researcher has to publish to gain visibility and reputation[36]. Unpublished findings do not exist for science (Havemann 2009, p. 8). In former days, publication was necessarily connected to "physical realization" or recordings (cf. Boyce and Kraft 1985, Broadus 1987), however, nowadays the spectrum of publication methods shows a higher variety. In particular, digital publication outlets were introduced, which also offer new means for analyses (number of accesses instead of citations, cf. Altmetrics[37]) and more importantly an easier and more timely dissemination of results. Furthermore, references are no longer restricted to written communications as recent studies have shown (Kousha and Thelwall 2012). The mere introduction of links as references provides new possibilities for evidence bases but also makes the set of references more unstable.[38]

In former years, repeatability and confirmability were major issues for bibliometric analyses. Since the accessibility to scientific publications was only possible in a physical (i.e. non-digital) form, bibliometric studies in the early years demanded a tedious setup, for which publications and the respective citations had to be collected manually. However, the digital age also brought a number of bibliographic databases which facilitated the monitoring of the increasing number of publications (cf.

---

35 A requirement that, in its full extent, is only completely satisfied with the introduction and spread of open access publications.

36 The genuine equivalent to monetary gain in the original production setting (cf. Chapter 2).

37 For a definition see http://altmetrics.org/manifesto/, last accessed on 2014/03/04.

38 I.e., if it is assumed that the written physical form at least in most cases ascertains the accessibility of a publication.

Borgman 1999, Borgman 2007, pp. 89f, Garfield 1970). Nowadays, bibliographic databases may already contain extensions for bibliometric analyses, like e.g. Scopus or WoS. Their coverage includes – but is not restricted to – all information concerning the references and citations of a publication, its author(s), its title and its source of publication.

Bibliometrics further covers the analysis of structural changes in research activity or orientation. For instance, research output is mapped or clustered and compared over a number of years (see e.g. Small 2006, Börner, Chen and Boyak 2003, Noyons, Moed and Luwel 1999, Noyons, Moed and van Raan 1999, Cahlik 2000). In this way, research topics that emerged in the analyzed time period can be traced back to their origin. Furthermore, changes in research topics, especially connections with other topics but also decline or a new orientation can be revealed.

In this thesis, bibliometric indicators are used to 1) organize the bibliometric data according to its topical focus and 2) identify those topics that can be called innovative (as assessed by bibliometric indicators). As explained in more detail in Section 3.2.1 (see also Section 2.2 above), citation analysis cannot be used as the necessary time lag prohibits the application to recent years. Still, the references of the analyzed documents can be used in order to give clues for the structure of the topic. This allows the computation of further indicators that are presented in this work.

3.1 Bibliometric Databases

As stated in Section 2.4, the bibliometric databases used in this work are Thomson Reuters' WoS and Elsevier's Scopus. Both databases collect information about scientific publications in various forms. Even though Thomson Reuters has a second database (Web of Knowledge) especially for patents, books and conference proceedings, the latter two may also be included in the WoS. As explained above (Section 2.4), the initial choice for a database can influence the results of an analysis. Furthermore, the outcomes are also highly dependent on various other decisions like document types, field delimitations, time windows etc. For instance, studies have shown that the in- or exclusion of conference proceedings can have major effects on the publication and citation counts for at least some fields, e.g. Computer Science (Bar-Ilan 2010, Michels and Fu 2014). Thus, in the following an overview of these influence factors in regard of the usage of bibliometric databases is given. First, however, the coverage is discussed.

A systematic comparison of both databases would demand the calculation of the overlap in documents in both databases. However, in order to measure this overlap, one would have to compare the textual information of all articles in both databases. Jacso (2005) compared the overlap in citing items in WoS and Scopus (and Google Scholar) and showed that there are not only differences in the coverage of publications but also – as a consequence – of citing items. Therefore, the number of citations an article receives might vary according to the database coverage. There is one article in his document set, namely "Persistence in nonequilibrium systems" by Maumdar, which is cited 82 times according to the WoS, but only 58 times in Scopus. A similar study was performed by Meho and Yang (2007). According to them, there is a total overlap of 58.2% of citing items in Scopus and WoS, but the number of unique citing items found in Scopus is slightly higher than in WoS (26.0% in comparison to 15.8%). The coverage of the items may also vary in the different fields. Rankings (in their case the

ranking of faculty members) according to citation numbers might change if the information provided in the WoS is supplemented with the information in Scopus. They show that mostly middle-ranked entities are affected by this.[39] Therefore, a complete picture might require the analyses of both databases.

A later study by Meho and Rogers (2008) shows the variations in bibliometric indicators regarding different databases for the case of Human-Computer Interaction researchers. As in the former study, they show that due to the low coverage of proceedings, researchers that focus their scientific communication on this publication type (e.g. Computer Scientists) are disadvantaged especially in the WoS. In their set of top 22 citing journals and 20 conference proceedings, the WoS only covers 19 journals and 8 conference proceedings, while Scopus contains all of them. Following their reasoning, it is thus necessary to use "Scopus instead of Web of Science for citation-based research and evaluation in [Human Computer Interaction]" (Meho and Rogers 2008, p. 1724). Therefore, in the course of this thesis, both databases were used. The selection depended on the document types and requirements for the specific tasks. In particular, for an analysis focussing on proceedings, Scopus was used (see Section 9.2.1).

Note that the findings above also show the necessity for an error-free document type assignment. Sigogneau (2000) identified 18 out of 111 Physics journals in the WoS, for which "more than 70% of papers were Proceedings in 1996" (Sigogneau 2000, p. 599). In a White Paper, Harzing explains the logic behind the categorization:[40] "presenting an early version [...] appears to mean that your paper is downgraded by ISI to be a "conference proceedings paper" even though the conference in question doesn't even publish proceedings". According to her, ISI/Thomson Reuters detects these publications via the acknowledgements. Since some bibliometric indicators, e.g. the Impact Factor (see Section 3.2.3), are restricted to certain document types, misclassified publications as well as their citations are wrongly excluded from the calculations. In the two examples provided by Harzing, the authors lose at least half of their articles due to such a misclassification.

Further important information might be the coverage of abstracts, for which Jacso (2005) states that it is slightly higher in Scopus than in WoS. Since some studies use keyword searches that also include the abstracts, a difference in coverage of abstracts might also affect the number of publications that can be found.

In a study on in- and exclusion of journals in the WoS, Michels and Schmoch (2012) were able to show that an observed growth in publication numbers also stems from the fact that the database coverage is extended. An increase in article numbers for the period 2000-2008 by 34% was calculated. But half of the associated journals, that lead to additional articles, are either journals that make a reappearance in the database after a long time or are newly introduced in the database even though they have been published for at least 10 years (i.e. have a volume number higher than 10). Thus, growth effects

---

39 In this context, Schneider (2012) also pointed out that the mere ranking of entities according to (bibliometric) indicators can hide the fact that the actual values only differ slightly.

40 "Working with ISI data: Beware of Categorisation Problems", Anne-Wil Harzing, http://www.harzing.com/ISI_categories.htm, last accessed on 2012/10/18.

should not be interpreted in absolute values but only in relation to other study subjects or the overall growth rate.

3.2 Bibliometric Indicators

With time, a number of standardized bibliometrical indicators have been introduced to measure the scientific performance or standing of journals, authors, organizations, regions or countries. As a consequence, several discussions dealt with the scientific and ethical implications of their application.

Ibáñez, Larrañaga and Bielza (2011) use a Bayesian Network to analyze the relationship between different bibliometric indicators. Even though their main goal was to provide means for the deduction of reasons of a journal's score in bibliometric indicators, it should be noted that these analyses also show that there might exist (latent) dependencies between indicators that also must be taken into account when they are used to evaluate a journal. Similar conclusions can be made in regard to the assessment of single documents, which corroborates the usage of multivariate analyses in this thesis (see Chapters 6 and 9).

Another factor that should be born in mind when applying bibliometric indicators is that the observed entities – scientists – are aware of the observation and the metrics used for assessment – bibliometric indicators. It is only natural that the study subjects adapt their behaviour (Bornmann 2011, Michels and Schmoch 2014). For instance, Fraser and Martin (2009) reported a change in usage of vocabulary or certain terms in publications to increase the chance for publication.

Apart from the most prominent ones, there have been a couple of indicators that were proposed to measure impact or other new values in a different way. For example, Leydesdorff (2007) introduces the betweenness centrality of journals as an indicator for their interdisciplinarity. Klavans and Boyack (2010) introduce the notion of a thought leader denoting a scientific actor (i.e. a university, state or nation) that "is building on the more recent discoveries in a field" (Klavans and Boyack 2010, p. 546). They criticize previous work that was restricted to identify leadership in single scientific fields by using journal based metrics. While their argument is valid in suggesting the usage of co-citation instead of journals to classify the leaderships, their definition of a thought leader is focused on the (re)use of already published ideas. A thought leader might per se not have recent work that he can refer to when conducting research at the research front. Thus, restricting leadership to those people that rely on recent findings of others might be counterproductive.

In the following, a variety of bibliometric indicators is discussed to show the possibilities and limitations of bibliometric analyses. First, citation analysis in general is discussed, then two derived indicators for the assessment of individual researchers or journals respectively are exemplarily shown. The latter of the two, the Journal Impact Factor, which is in more detail explained later (see Section 3.2.3), is also applied in the rule-based part of the approach.

3.2.1 Citation Analysis

Citations

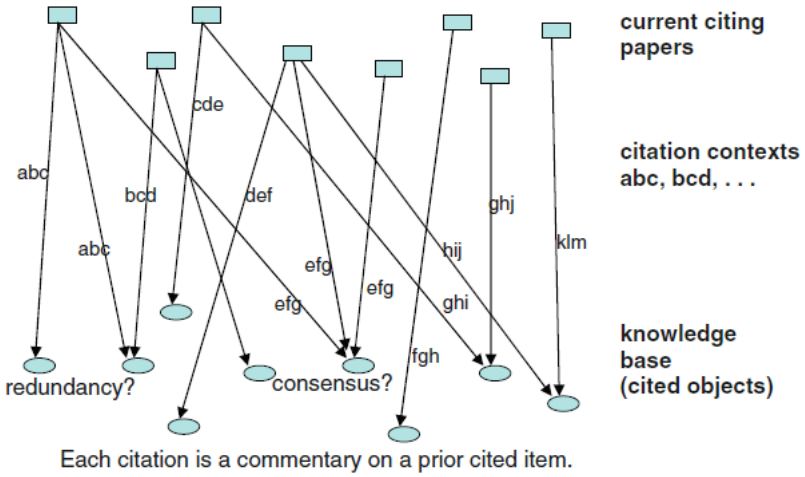
A very crucial part of bibliometric analyses is the assessment of individual scientists or institutions. Most assessments are based on indicators that use citations. As already hinted at above (cf. Sections 2.2), the number of citations for a publication or an author might be influenced by various factors. For instance, a Matthew effect (Merton 1968) can be observed, i.e. authors with highly cited papers tend to gain more citations than comparable scientists (cf. definition on p. 24). Also, as was shown by Schubert and Michels (2013), MacRoberts and MacRoberts (1989) and Lancaster, Lee and Diluvio (1990), the publication source and its location can affect the number of citations. It has to be considered that citations do only partly reflect the classification or surrounding of a publication. A huge part of the citation process is also politics and who-knows-who (see e.g. Bornmann et al. 2012).

MacRoberts and MacRoberts (1989) name different problems of citation analysis, among them biased citations, self-citations and the different types of the citations. The notion "biased citations" means on one hand that former works are disproportionately used or cited. On the other hand, sometimes only the secondary sources are cited, so that the "'credit' given was taken from the discoverer and allotted to someone who had nothing to do with the discovery" (MacRoberts and MacRoberts 1989, p. 344). Different types of citations refer to the citation context and the usage of citations. As will be discussed later in more detail, a citation does not necessarily imply agreement.

Cronin (1984, pp. 35ff) gives an overview over different classification systems for citations. Independently of the depth of the classification scheme, redundant and perfunctory citations can be distinguished. Also, differences can be observed on the level of reference, as citations can refer to similar applications or datasets (or even a reuse thereof), to a similar theory or methodology, to results or statements corroborating the findings in the citing paper, to substantial background or methodologies or to findings or statements that are criticized, refuted or doubted. With the exception of the reuse of a dataset, all types of citations indicate related topics in citing and cited paper. Also, Cronin points to the fact that citations are sometimes used excessively and unnecessarily (cf. "obsolete" or "self-serving" citations in Thorne 1977, cited after Cronin 1984, p. 64). Liu (1993) and MacRoberts and MacRoberts (1996) conducted related studies for the context of citations.

Small (2011) states that "within a specialty the researchers are more likely to be using concepts, data, tools, and methods in an instrumental manner" (Small 2011, p. 383), i.e. cited work forms a basis for further improvement while avoiding the reinvention of the wheel (cf. Section 1.1). This assumption justifies the usage of the reference list of the observed documents to establish connections between documents and topics for this thesis (cf. Section 5.2).

Figure 13: Use of citations.



Source: Small (2011, p. 374)

Small (2011) investigated the contexts of citations and their respective sentiment to reveal their intentions (Small 2011, see Figure 13). Table 3 shows the model behind his work, which includes the citation context, i.e. the research topic, as well as the commentary or sentiment, which might be one out of the seven listed in Table 3.

Table 3: Cue words for the usage of citations.

Sentiment	Sample cue words
Importance	Significant, best, crucial, fundamental, ideal, notable, remarkable ...
Utility	Employed, with, applied, used, using, utilizing, application ...
Report	Described, discussed, account, stated, published, reviewed, observed ...
Consensus	All, common, majority, most, typical, widely, well-known ...
Uncertainty	May, might, could, should, possible, potentially, projected ...
Differentiation	Contrast, differs, difference, compared ...
Negation	Not, although, however, but, failed, controversial ...

Source: Small (2011, p. 379)

As Garfield (1974-76) puts it, “any paper cited ten times in one year is *ipso facto* significant [...,] a paper cited ten times in each of two successive years is well on its way to citation stardom” (Garfield 1974-76, p. 419). This assumption led to the now wide-spread application of citations as a measure for importance, significance or popularity. On the other hand, also less qualitative work might result in citation scores. Opthof (1997) describes this very illustrative in his paper about the JIF (see Section 3.2.3) as an impact metric:

“It is obvious that citations like “we confirmed previous data of Opthof et al.....” and “by misinterpretation of their own data Opthof et al. erroneously suggest that.....” or “the fraudulous work of Opthof has retarded the field of autonomic influences on heart rate for decades” constitute different qualifications even if they all are scored as one citation” (Opthof 1997, pp. 1f).

Another factor with possible discrepancies is the number of citations a document received. Citations for one document are calculated by counting the number of documents that reference to it. As already discussed above (see Section 3.1), the citation count might vary depending on the coverage of the respective citing documents. The number of citations for a document can only be calculated correctly, if all documents that cite the respective document are included in the database. Garfield (1974-76) hints that the list of references might influence the citation rate and thus be an indicator for a papers' quality. The accuracy and coverage of the topic by a papers' reference list might also indicate the diligence and the effort the authors put into the whole work. Thus, this can be seen as one indicator for the overall quality. Nonetheless, it is a difficult indicator for the innovativeness of a publication. In this thesis, alternative features of a reference list are tested as an indicator for the innovativeness of a publication.

Bornmann et al. (2010) found that papers with extraordinary impact mostly contain references to other highly cited papers. Thus, the notion "on the shoulders of giants" was corroborated at least for papers with high citations in all fields.

Interpretation

Many studies rely on citations to measure innovativeness and impact, since a high innovativeness and impact might reflect in high citation rates. Still, a low citation rate in turn does not necessarily indicate a paper of low innovativeness or importance. One might argue that sooner or later the scientific community would discover and cite any paper of importance, but as the introductory example of Mendel showed, the induced delay can last for decades.

It seems that all bibliometric studies that rely on citations assume that the scientific community is infallible and will highly cite everything that is important and ignore a paper if and only if it is unimportant. In the case of Mendel and others the scientific community failed and there is no reason why it can be assumed that this was a single case (cf. Garfield 1970, Hook 2002, Ohba and Nakao 2012, Soh et al. 2012, Glänzel, Schlemmer and Thijs 2003, Kozak 2013). Also, only those examples are known where the scientific community neglected to acknowledge a paper for which this implicit decision against the paper was redeemed later (cf. Chapter 2.2). However, various (yet unknown) cases might exist where scientific findings are implicitly rejected despite their importance and innovativeness.

High citation counts might not only indicate a high importance, but a high ambiguity or maybe also fundamental research. This does not imply that fundamental research is less important, but per definition the citedness of such work is higher than in other work (Boyack et al. 2013). Also, fundamental research can be cited more often because it applies in various contexts. The question remains if a fundamental work can be used as a basis for an emerging topic, i.e. if a new topic starts with fundamental research or if it builds the basis thereof.

Also citation counts are in most cases independent of the context of a citation. Thus, if a citation is used to denounce another work, it is counted anyway. As argued above, the context, e.g. "Our results suggest that the work done by ... does not hold in realistic conditions", is not considered in the citation analysis (see Opthof 1997, Cronin 1984, pp. 35ff).

Adams (2005) suggests that early citation counts reflect the overall or later citation counts. A closer look at the applicability of this observation on new topics is taken in Chapter 8. The time lag between the publication of a document and its citations makes the citation analysis useless for the assessment of current events. Nonetheless, the references of a document can give hints for its character, innovativeness and especially topical relatedness (cf. p. 15). For the debate on the appropriate citation window in which citation rates are calculated, the study mentioned above by Adams (2005) shows that the correlation is higher between citations in the first two years with the total number of citations than with later citations. Thus, an article that receives many citations in the beginning can be expected to have a high number of total citations. But the implications on its later citation rate are not as confident. Therefore, smaller citation windows might be an option to study the citation rate of an aggregation, but not to detect single high impact papers since for these the citation rate might differ from the usual pattern. Adams (2005) also confirms that the correlation between early and later/total citations also depends on the field of science.

The number of references in an article may vary depending on different factors, most prominently the scientific discipline. According to Garfield (1974-76), shorter reference lists might derive from the fact that a field is more specialized than others or because a different literary style is used. The former case also applies to new or small topics.[41]

According to Zuccala (2012), in order to measure scientific impact with citations, all sources should be used, i.e. articles as well as notes, letters etc. It is important to acknowledge different citation behaviour in different fields (Harwood 2008) and research studies (Hewings, Lillis and Vladimirou 2010). The time lag as well as the citation counts are dependent on the disciplines. In particular, citations are more common in some disciplines than in others. Thus the citation count of an institution or country might depend on its disciplinary portfolio.

Further differences in citation behaviour might arise due to varying types of science (e.g. empirical versus conceptual), personal concepts of appropriate citing and the publication outlet (Harwood 2008). A further study in agricultural research also showed that meta-level concerns could influence the decision to cite a document (White and Wang 1997).

Levitt and Thelwall (2008) extended the idea behind disparating citation rates for different fields by an analysis of citation rates for monodisciplinary vs. multidisciplinary articles. The surprising result was that monodisciplinary articles were even higher cited than multidisciplinary ones. In the fields Life Sciences, Health Sciences and Physical Science, the citation rate was on average twice as high for monodisciplinary articles.

Citation Analysis Manipulation

Citing previous work can be a way to refer to steps and conclusions already made in a scientific elaboration. In Sections 3.2.2 and 0, concrete examples for bibliometric indicators and possible manipulations are given, which in the majority rely on artificially increasing the citation rate of certain publica-

41 Garfield (1974-76) uses the term “field” in his explanation but the same holds in this context for topics.

tions. Also, Tagliacozzo (1977) showed that all scientists seem to have a tendency to cite their own work more often than that of colleagues. This tendency is not correlated with the "number of citing co-authors, size of bibliography, and author productivity" (Tagliacozzo 1977, p. 264) so that Tagliacozzo suggests that the degree of self-citation might simply be a personal trait.

Because of that, some institutions have decided to calculate bibliometric indicators based on citations under the exclusion of self-citations (instead or additionally). The additional cost is relatively low, since these self-citations can be detected by a mere comparison of the authors of the citing and the cited work (cf. Aksnes 2003). Glänzel and Thijs (2004) emphasize that – due to potential biases caused by author-mismatches – self-citation analysis should always only complement any other kind of bibliometrical analysis but not substitute it.

Another form of self-citations, journal self-citations (cf. p. 53), is more difficult to detect. Since a set of bibliometric indicators assesses the impact of a journal by the number of citations to its articles, some editors urge the authors to include more citations to papers previously published in the same journal, if they want their article to be published.[42] Of course, these citations could be excluded similarly to the author self-citations, but the impact would be higher in total and less accurate. The set of journals that is important for a single topic might be very restricted and thus, excluding all journal self-citations might hurt the evaluation of more specialized work.

Another way to push citation counts for one's own work is the so called "Salami Publishing" (Abraham 2000, Roberts 2009). In this case, authors deliberately publish smaller portions of their work in a larger number of publications. In this way, there is (seemingly) more output, more stuffing for the CV and more publicity and thus also a higher chance of being cited. According to Leimu and Koricheva (2005b) this strategy does not necessarily result in higher citation rates, though.

Sleeping Beauties and Similar

Some scientific publications do not receive attention or citations at all for a very long time. If it is discovered later-on that this was unjustified and that they indeed deal with a topic of uttermost importance, they are called "Sleeping Beauties" (van Raan 2004). Normally, a "prince" comes along in the form of another publication that cites the "Sleeping Beauty" for the first time (see van Raan 2004). The attention of the scientific community is then attracted and citations are emitted to the "prince" and/or the initial paper and the topic is finally picked up in the scientific discourse. The phenomenon itself is also called the Mendel Syndrome for the reasons discussed above (van Raan 2004).

Contrasts to Sleeping Beauties are publications that are affected by the Matthew Effect (cf. definition on p. 24) or by "obliteration" (Merton 1968, Garfield 1979). The Matthew effect describes the phenomenon that "For to all those who have, more will be given", i.e. already famous scientists receive more attention and also more citations than their colleagues (Wang, Yu and Yu 2011). The original paper by Merton (1968) also mentions, that this is not only true when considering a collection of publications from various authors with different reputation level but also when dealing with one paper by

42 http://scholarlykitchen.sspnet.org/2012/04/10/emergence-of-a-citation-cartel, last accessed 2012/08/29.

a multiple number of authors. To stress this point, Merton cites a laureate in Chemistry "When people see my name on a paper, they are apt to remember *it* and not to remember the other names" (cited after Merton 1968, p. 57).

Note that this effect deals on the author level, while the "obliteration" concerns the publication level, as does the Sleeping Beauty. If a publication "becomes so generic to the field, so integrated into its body of knowledge" (Garfield 1979, p. 365), it might be cited less explicitly because it is already considered as common knowledge. In both cases, as well as for the Sleeping Beauties, citation analysis might fail to measure the impact of certain publications.

In summary, the motives for the distinct status of the Sleeping Beauty paper can be manifold, including the following:

- The importance of the claims or implications of the paper are not visible or not acknowledged by the scientific community
- There is no suitable application, yet, for the claims and implications of the paper
- The output source, journal etc. has a very low visibility

Independently of the reasons for the "sleep", as soon as the paper is rediscovered, it is clear that it plays an important role for the scientific discourse (otherwise it would not have been discovered at all). Thus, its citations and attention at this time are even extraordinary. Furthermore, a variant of the Matthew effect occurs in this context as well: The more attention the paper gets, the more often it is cited. The more citations it receives, the more attention it gets, etc. Thus, an extensive neglect is compensated by a disproportional amount of attention afterwards.

As Braun, Glänzel and Schubert (2010) point out, the study of citation behaviour for Sleeping Beauties and princes helps to understand the information flow in scientific publications induced by citations. More importantly for this thesis, though, is the conclusion that citations are a rather unreliable metric and indicator for the importance of a paper. This notion is discussed in more detail in Chapter 8. However, Sleeping Beauties enable the analyses of papers that seemingly introduce a topic for which the scientific community is not ready yet (cf. the "adjacent possible" in terms of innovation as explained in Section 2.1).

As already stated above, citation rates cannot be used as an indicator for a new topic without a time-lag. The concept of Sleeping Beauties even corroborates this statement. However, Burrell (2002) proposed a formula to calculate the chance for citation of a yet uncited paper. In his model, the probability diminishes over time, thus, with every year that a paper remains uncited, the probability that it ever will be cited decreases. This is a clear contradiction to any Sleeping Beauty that has been awakened, but one has to bear in mind that Sleeping Beauties are per se exceptions from the rule.

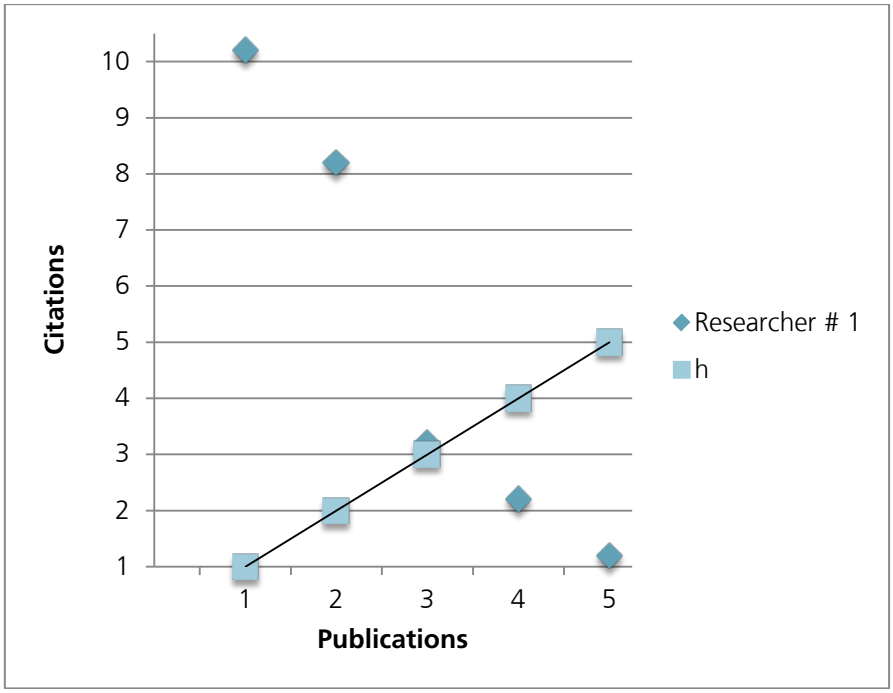
Costas, Leeuwen and Raan (2011) also distinguished between different types of citation behaviour to identify cases of Mendel syndromes or Sleeping Beauties, namely normal papers, delayed papers and flashes in the pan. The measurement of citations was based on a comparison with the field average values. Concretely, they compared the point in time when a publication has received 50% of its citations. If this happens before 75% of the publications in its field have received the same proportion of their citations, it is called a flash in the pan. If the "break-even point" is reached afterwards, it is la-

belled as a delayed publication. Thus, their definition of Sleeping Beauties somewhat differs from that given by van Raan (2004), as they focus on the delay of citations for a publication without making the high citedness of this publication a necessary condition.

3.2.2 h-Index

The h-index (Hirsch 2005) was introduced to measure the performance of individual researchers (see for example Cronin and Meho 2006). It measures the maximum number of publications *h* of a scientist that in turn also have at least *h* citations each. As a way to facilitate and illustrate the computation of the h-index, the publications are first ordered according to their number of citations. Figure 14 shows this for a fictitious distribution of publications and citations. The publication with the highest number of citations is at the left hand side. All publications are ordered so that the citation number is diminishing from left to right. As the h-index denotes a point wherein the number of publications and citations are equal, it can be found via the intersection of the publication points and the diagonal. In Figure 14, this corresponds to an h-index of 3 as the third publication with 3 citations is the last above the diagonal. In this particular case it is also the intersection point.

Figure 14: Exemplary calculation of the h-index for a researcher with 5 publications, which have been cited 10, 8, 3, 2 and 1 time(s) respectively.



Source: Own illustration

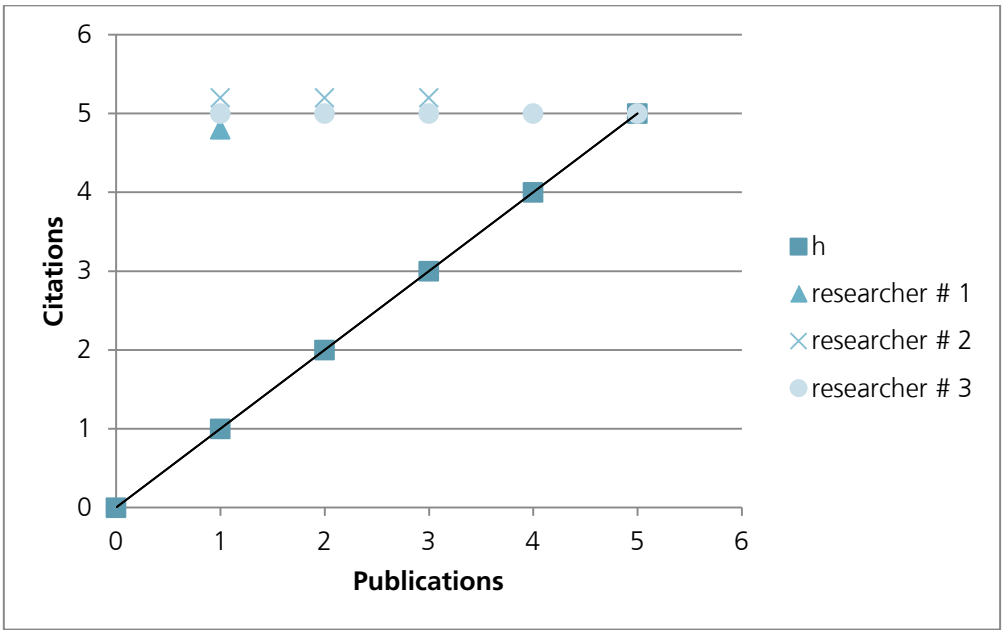
Notes: For a better readability the citation points are relocated with a small deviation from the horizontal lines.

The h-index has been criticized, even though no major flaws due to missing publications of individual researchers could be shown (Rousseau 2007). However, the h-index favours older researchers since 1) “well-established researchers and projects are cited disproportionately more often than those that are less widely known” (Wendl 2007, p. 403), 2) “scientists who have worked in their field for a longer period than younger scientists have a greater chance to publish more” (Bohlen und Halbach 2011,

p. 193, on a similar notion see also Vinkler 2007) and 3) “the h-index increases with age” (Kelly and Jennions 2007, p. 403).

As a simple example, a set of researchers is analyzed who all publish one publication per year that receives 5 citations. The only difference is the time span in which they have been publishing. Figure 15 shows the respective illustration for three researchers with this publication pattern. Researcher # 1 has been publishing for 1 year, while researcher # 3 has been publishing for 5 years and thus has already 5 publications with 5 citations each. In general, a researcher with this pattern who has been active for Y years (with $Y \leq 5$) will have an h-index of Y.

Figure 15: The h-index for three exemplary researchers who publish annually one publication, which in turn is cited 5 times.



Source: Own illustration

Notes: Researcher # 1 has been publishing for 1 year, researcher # 2 for 3 years and researcher # 3 for 5 years.

Thus, even though the citation rate is equal for all researchers, the “older” researcher profits from this calculation. Also, if one researcher is added who publishes a paper that receives 10 citations, he will have the worst h-index value of 1. And if the example is extended even further so that it deviates from the so far static citation rate, it becomes apparent that an older researcher receives more citations simply due the fact that his papers have been on the market for a longer time.

Therefore, a variety of alternative calculations for the h-index have been proposed, like the a-, m-, r- and ar-index, that bring their own advantages and drawbacks (see Table 4 for a comparison, Thompson, Callen and Nahata 2009).

Table 4: The h-Index and alternative metrics.

Metric	Definition	Advantages	Disadvantages
h-Index	A scientist has index *h* if *h* of his papers have at least *h* citations (and the other papers have less than *h* citations)	Combines qualitative and quantitative features Identifies a core of high performance articles (Hirsch core *h*)	Insensitive to highly cited work Difficult to compare across disciplines
a-index[43]	The average number of citations in the Hirsch core *h*	Takes into account the skewed nature of citations	Focuses more on the impact of the Hirsch core *h* instead of its size Can be sensitive to few highly cited papers
m-index[44]	The median number of citations in the Hirsch core *h*	Takes into account the skewed nature of citations	Focuses more on the impact of the Hirsch core *h* instead of its size
r-index[45]	The sum of all citations in the Hirsch core *h*	Does not punish a high variety of citation counts for the publications in the Hirsch core *h*	Focuses more on the impact of the Hirsch core *h* instead of its size Can be sensitive to few highly cited papers
ar-index[45]	r-index normalized by the years publishing	Takes into account the age of the researcher	Like r-index

Source: Adapted from Thompson, Callen and Nahata (2009)

Bohlen und Halbach (2011) test the suitability of the h-index and comparable indicators for a preselection of job application candidates but could not find any better way than "the old-fashioned (and time-consuming) method of reading the papers" (Bohlen und Halbach 2011, p. 196). There were also studies suggesting that even though the h-index could predict the future success of individual scientists (Hirsch 2007), it is not an indicator for a single article's future citation rate (Hönekopp and Kleber 2008).

The h-index exemplarily illustrates the problems associated with bibliometric indicators and also citation analysis. In particular, various factors can influence the recognition of a work and interpretations of the indicators should always consider their limitations. In most cases, only a multivariate analysis can take into account the volatile nature of bibliometric indicators. More details are given in Chapters 6 and 8, wherein such analyses are performed.

43 Bornmann, Mutz and Daniel (2008).

44 Hirsch (2005).

45 Jin et al. (2007).

3.2.3 Journal Impact Factor (JIF)

Among the journal-related indicators, the Journal Impact Factor (JIF) is probably one of the most prominent ones. This indicator measures the impact of a journal based on the citations for its articles. The exact formula is (Moed and van Leeuwen 1995):

$$JIF(y) = \frac{Cit_y(art,not,rev,let)}{Pub_{[y-2,y-1]}(art,not,rev,let)} \quad (3)$$

where $Pub_{[y-2,y-1]}(art,not,rev,let)$ are all publications of the type article, notes, reviews and letters in the specific journal in the years *y*-2 and *y-1* and $Cit_y(art,not,rev,let)$ are all citations in the year *y* to the publications in the denominator.

This indicator as well has been in life-long critic for biases due to the field and the language of a journal (Seglen 1997). Furthermore, Thomson Reuters uses all citations to all articles of a specific journal and divides this number by the number of "citable" articles in the journal. In the formula given above, it seems that citable clearly refers to all documents of the named types. However, reverse engineering suggest that other document types have been used by Thomson Reuters for the calculation of the JIF (Moed and van Leeuwen 1995). Not only did these types not match the ones given above, they also differed for the calculation of numerator and denominator.

All in all, the exact definition of "citable" is not accessible and therefore, it can only be assumed, which formula is used (Moed and van Leeuwen 1995); "During the course of our discussions with Thompson [sic] Scientific, PLoS Medicines potential impact factor – *based on the same articles published in the same year* – seesawed between as much as 11 (when only research articles are entered into the denominator) to less than 3 (when almost all article types in the magazine section are included, as Thomson Scientific had initially done – wrongly, we argued, when comparing such article types with comparable ones published by other medical journals)" (The PLoS Medicine Editors 2006, pp. 707f).

As is the case with nearly every bibliometric indicator, the JIF is sometimes applied in a way that was not intended by the inventors, e.g. to measure the performance of individual scientists or groups of scientists. The JIF was also suggested as a predictor of an article's future citations (Hönekopp and Kleber 2008, Ogden and Bartley 2008). Opthof (1997) analyzed the applicability of the JIF for the assessment of the performance of single publications, researchers and group of scientists. At least for the latter two he was able to show that the JIF cannot be used in such a context. Regarding the results by Chang, McAleer and Oxley (2011) for the correlation between a JIF for 2 and 5 years and with and without self-citations, one can at least deduce that this indicator is stable independently of the considered time span and the inclusion of (journal-) self-citations. Nonetheless, with regard to this thesis, it can be concluded that the JIF is no suitable metric to detect high-impact articles, since as Chang, McAleer and Oxley (2010) already stated "Great papers appear in great journals [and] [a]ll great journals publish great papers [but] [n]ot all papers in great journals are great." (Chang, McAleer and Oxley 2010, p. 3, Chang, McAleer and Oxley 2011, p. 19). This has also already been corroborated by the findings by Leimu and Koricheva (2005b): "publication in a prestigious [high JIF] journal does not by itself guarantee high citation rates" (Leimu and Koricheva 2005b, p. 32). Still, for the contributions of

this thesis, the JIF is applied to measure the reputation of a journal in order to detect patterns for emerging topics (cf. Chapter 6 and 9).

There have been suggestions for a topic-adjusted JIF (Takahashi, Aw and Koh 1999, Uehara et al. 2003). In this adjustment, the JIF only takes into account the publications and the citations from a specific topic. Fu et al. (2011) showed that there can be major differences between both calculation methods and thus both metrics should be taken into account when assessing a journal with mixed topic focus.

A recent debate about "Honorary Authorship" in the journal Science shows the simplest form of indicator manipulation: The number of publications and thus citations can be increased for both the real and the honorary authors. So, the incentive for this behaviour is two-fold, as "researchers add the names of prominent scientists to boost their papers' credibility, and senior scientists demand that their names be added to the work of younger researchers" (Honorary Authorship 2012, p. 1453). Still, post-detection is difficult and thus only "concerted efforts by institutions, authors, and journals are needed to put an end to this fraudulent and unethical practice" (Greenland and Fontanarosa 2012, p. 1019).

On the journal side of publications and performance measurement, there have also been reports of so-called citation cartels. In a first iteration of pushing their JIF, journals cited their own articles extensively, or rather, they let them be cited. Self-citations of journals are easily to detect and some journals have already been excluded from the Journal Citation Report (JCR) by Thomson Reuters for exactly that reason.[46] Thus, in a second iteration, journal editors started to collaboratively push their citations by urging authors in one journal to include more citations for their other journal.[47] There have been reports of editor responses that included statements like the following: "you cite *Leukemia* [once in 42 references]. Consequently, we kindly ask you to add references of articles published in *Leukemia* to your present article" (Wilhite and Fong 2012, p. 542). These – unsubtle – hints clearly are intended to boost the citation rate of journals, with which the editor might be also directly or indirectly connected. The authors on the other hand "are rewarded for acquiescing because their manuscript is published" (Wilhite and Fong 2012, p. 543). If more than one journal is involved as in the example above, citation manipulation is not that easily detected. Calero-Medina and Costas show a method to detect such "citation cartels" by looking for common editors and conspicuous citation patterns.[48]

3.3 Maps of Science

There have been different approaches to visualize and classify scientific publications according to topical relations. These maps can be used to analyze the interaction of different scientific topics in a

---

46 http://scholarlykitchen.sspnet.org/2011/10/17/gaming-the-impact-factor-puts-journal-in-time-out/, last accessed on 2012/10/25.

47 http://scholarlykitchen.sspnet.org/2012/02/02/when-journal-editors-coerce-authors-to-self-cite/ http://scholarlykitchen.sspnet.org/2012/04/10/emergence-of-a-citation-cartel/, last accessed on 2012/10/25.

48 "Journal Citation Cartels: Can they be detected?", Clara Calero-Medina and Rodrigo Costas, http://www.helsinki.fi/kirjasto_old/keskusta/images/verkkari/Calero-medina%20costas%20121011_Cartels_presentation_Helsinki.pdf, last accessed on 2012/10/25.

fixed time period or their development over time. Thus, they serve as an illustration for later manual interpretation that highly depends on the primarily chosen set of indicators used to map the scientific landscape in a 2-dimensional space. In most cases, only one feature is selected to cluster and locate the documents, so that all interpretations of a map are basically restricted to its expressiveness as well.

As will be shown in Sections 3.3.2 and 3.3.3, most mapping algorithms use co-citation or co-word-analysis to cluster the documents. Therefore, a two-dimensionality of science is assumed (cf. Griffith et al. 1974, p. 363 with regard to the work by Price 1966). Other analyses have built upon this assumption and used single-featured clustering for this purpose (see e.g. Small and Griffith 1974, Small, Sweeney and Greenlee 1985, Small and Garfield 1985), but also resulted in maps that combine different features (Ahlgren and Colliander 2009, Janssens, Glänzel and Moor 2008). In these studies, the purpose was solely to create a map to illustrate, analyze and discuss ongoing trends and developments in science. However, as was already mentioned, the dimensional reduction to one or few features might influence the view so that the selection is crucial. Or as Leydesdorff (1987) already wrote, the so-gained results "are largely dependent on the choice of options offered by the computer programme [...so that] there remains always the technical question of how the lines drawn between the different points can be legitimized" (Leydesdorff 1987, p. 321).

It should not go unmentioned that Klavans and Boyack (2009) also provided a study in which a consensus map was built using the overlap of other maps. This approach sidesteps the problematic preselection of features but also has to deal with combined restrictions with regard to coverage.

To better interpret the implications of these features, comparisons of different clustering approaches have been performed (Cahlik 2000, Börner, Chen and Boyak 2003, Rafols and Leydesdorff 2009, Klavans and Boyack 2006a, Klavans and Boyack 2006b). For these purposes, metrics like the following were available:

- Coverage: How many publications are included in the map
- Accuracy: How many publications are correctly aggregated in a cluster
- Dispersion of publications: How many publications are covered by more than one class and if so, by how many
- Bias: Are there any fields or topics that are preferred by the clustering algorithm and thus represented unproportionally

This thesis uses variations of the former two metrics, namely Recall and Precision, as introduced later (Section 4.1.3). Similar to the work above, the calculation of such metrics facilitates the comparison of different mapping approaches.

The following subsections list different possibilities for the mapping of scientific activity, which can be compared to the clustering approach presented in this thesis. The collection starts with the concept of classifications, which denote in contrast to the other elements not methodologies but structures. Classifications can be constructed via the methods listed below, but are usually characterized by a static assignment to manually labelled categories. Sections 3.3.2 to 3.3.4 describe citation-based mapping techniques. Author and term-based networks follow, before the selection of research fronts in scientific databases is described in more detail.

3.3.1 Classifications

Publications from different research fields can be distinguished with the help of article or journal classifications. The bibliometric databases used in this thesis provide such classifications based on the field of research of the publications or their respective publication source. Further categorizations can be added to these databases as well.

The associated classification for the WoS is the "ISI Subject Categories". In this classification, one or more categories are assigned to each journal. The drawbacks and advantages of an ambiguous journal classification are discussed below. As a univocal alternative the classification for the Essential Science Indicators assigns only one out of 22 categories per journal[49], but is restricted to the defined journal set[50]. Scopus also comes with an already implemented journal classification system. In this classification, multiple assignments of journals and thus publications are also common but not as often the case as in the WoS classification. The classification system of Scopus is hierarchical in the sense that it has two structural levels, while the WoS categorization is flat. Counting all classes in the Scopus system (top and bottom level alike), 334 classes are distinguishable in total. By contrast, in WoS only 242 categories are used.

An ambiguous (journal or article) classification as implemented in both databases has its advantages and disadvantages. On the one hand, multiple assignments of categories have the capability for more specific classifications. If a journal or article in fact covers two or more scientific topics or disciplines, a distinct classification offers no possibility to represent this. In this case, information is lost. Furthermore, multiple classes show links between scientific areas more explicitly. The multidisciplinary journals and articles can be easily extracted from the data. In a 1:1-assignment, this has to be derived from additional data, e.g. works cited or the authors' affiliations.

On the other hand, multiple classes can blur the boundaries between the areas. In particular, since no detailed information is available, all classes are assigned with an equal weight – a notion, which in most cases is not correct since there is a dominant topic from which the research arises that uses methods, implications, etc. from other topics. In the WoS, there are up to 10 classes assigned to each journal. No information is available about the distribution of these classes on the articles. Thus, when using this information, it has to be assumed that all classes are equally represented in the journal. But such an equal weighting of classes can distort the results of bibliometric indicators.

For instance, in Scopus and the WoS, the number of classes for a journal averages 3.5 or 1.9 respectively. Publications in journals with multiple classes are also counted multiple times when the indicators for e.g. a country are calculated for these different fields.[51] By publishing one article in a journal with 2 classes, the publication count increases by one for both classes.

---

49 http://incites-help.isiknowledge.com/incites_19_live/appendixGroup/subjectAreaSchemes/essentialScienceIndicators.html, last accessed on 2012/11/09.

50 http://archive.sciencewatch.com/about/met/journallist/, last accessed on 2012/11/09.

51 An alternative would be the normalization with fractional counts for the fields, which similarly underrates publications if too many fields are assigned.

Beside manual classification, there have been various methods of clustering/categorizing scientific publications; Janssens et al. (2006) use a Latent Semantic Analysis on the texts of publications to cluster them accordingly. Rafols and Leydesdorff (2009) compare different kinds of scientific classifications and conclude that even though the mapping of the journals to the classifications varies, the classifications show a high similarity on an aggregated level. Manual efforts (e.g. Glänzel 2003) and semi-automatic or automatic approaches (Boyack and Klavans 2010, Sebastiani 2002, Small 1985) have been applied as well.

In the following subsections, different features are discussed which can be used to group and categorize scientific publications. These features highly depend on the citation data provided for the publications. Other methods to group scientific publications by their topical connection have been mentioned earlier – the only difference is that those maps are in most cases not labelled in such a way that classes can be derived. This thesis exploits the possibilities of an classification that assigns multiple classes to single documents or journals. With such a classification, the interdisciplinarity can be measured (cf. Chapter 7), which is implemented in some indicators of the system in this thesis (cf. Section 9.4).

3.3.2 Citation Networks

One possibility to group scientific publications is to generate a citation network on a set of publications. In order to do this, a directed graph is built based on the citation flow in the set of publications. The citation network allows deductions for the general structure of a topic or subtopics in a topic (i.e. larger clusters in the citation network) or its development over time (see e.g. Hummon and Doreian 1989, Kajikawa and Takeda 2009, Verspagen 2007).

The use of citations as a metric for topical relationship has been investigated in many publications. One of the first works dealing with this application in bibliometrics was written by Pinski and Narin (1976). Based on this paper, the famous Google PageRank algorithm was developed later, which measures the relationship between websites by links. In theory, using links between web pages is a mere continuation of the concept of citations between scientific works.

Rosvall and Bergstrom (2008) created a map of science as a network of scientific areas (modules) where the edges between the nodes, i.e. the modules, represent the citation flow between them. The size of the nodes indicated the citation flow within the specific module. The graph was designed bidirectional, i.e. edges were constructed in both directions from all nodes. The map reveals that applied science areas cite basic science extensively but not vice versa. The scientific disciplines are all connected to one another via citations either directly or indirectly, i.e. by other disciplines bridging a gap. Thus, they state that the structure of science rather resembles a “U” than a ring, where social sciences and engineering would be at the terminal ends. These disciplines are merely connected by a “backbone of medicine, molecular biology, chemistry, and physics” (Rosvall and Bergstrom 2008, p. 6).

Mina et al. (2007) identify the main path in the citation network of the most highly cited publications for the topic of “coronary angioplasty medical research”. The main path reflects and summarizes most efficiently the main research questions during the development of the topic over time.

Also, citation networks have been used to identify research fronts in a set of publications (Kajikawa and Takeda 2009, Shibata et al. 2009c, Shibata et al. 2009b, Winterhager and Schwechheimer 2002). In this case, publications are clustered according to similar citations and the most recent clusters can be labelled research fronts. Research fronts are explained in more detail in Section 3.3.7.

As a last example for the usage of citation networks, Huang and Chang (2011) analyzed the interdisciplinarity of a topic by the disciplines of the cited publications, but also stated that the ongoing trend of interdisciplinarity was more apparent in the co-author network, which will be discussed in Section 3.3.5.

3.3.3 Bibliographic Coupling & Co-citation

Co-citation analysis is the simplest form of citation analysis and is used to detect similar publications by comparing their sets of citing publications. One of the first implementations of a clustering approach for scientific publications based on co-citation was performed by Small, Sweeney and Greenlee (1985). The standard calculation of the similarity of two documents d_1 and d_2 corresponded to the absolute number of documents that cited both documents d_1 and d_2. A clustering then was applied using this absolute value. A hierarchical clustering, for instance, could start with the maximum value of the similarities of all documents in the set and aggregate them in a cluster, then take the pair with the next highest similarity and so forth. In the work by Small, Sweeney and Greenlee (1985), each citation for a document d_1 was normalized with the total number of citations from the emitting document d_{e1}, thus

$$cit(d_{e1}, d_1) = \frac{1}{\#\ documents\ cited\ by\ d_{e1}} \quad (4)$$

This normalization was only used for the selection of documents for which the number of citations exceeded a certain threshold in order to select a set of documents on which the approach was applied.

Furthermore, the co-citation of d_1 and d_2 was normalized with the number of citations both documents received:

$$\frac{\#\ co{-}citations\ of\ d_1\ and\ d_2}{\sqrt{(\#citations\ for\ d_1 * \#citations\ for\ d_2)}} = \frac{\sum_e (cit(d_{e1}, d_1) | cit(d_{e2}, d_2) > 0)}{\sqrt{\sum_e (cit(d_{e1}, d_1)) * \sum_e (cit(d_{e2}, d_2))}} \quad (5)$$

Based on this Small, Sweeney and Greenlee (1985) applied a hierarchical document clustering. They compared the results obtained for a fractional and an absolute threshold for document selection as well as those for a constant versus a variable co-citation clustering threshold. They showed that the bias caused by more "citation-intensive" fields like in Biomedicine can be reduced by the fractional citation counting method and thus would be a first step towards a consistent mapping of science.

As described above, Klavans and Boyack (2010) used co-citations to cluster publications independently of their original classification. Furthermore, Boyack and Klavans (2010) compared different clustering methods to identify the most suitable approach to represent research fronts. Bibliographic coupling uses the number of shared references, not citing documents, to calculate the similarity between two publications (see e.g. Sharabchiev 1989). According to the study by Boyack and Klavans (2010), bibliographic coupling outperforms co-citation analysis and direct citations. They themselves

showed differences to comparable studies in which direct citations sometimes provided better results, but in all studies, co-citation analysis performed worse than bibliographic coupling. Thus, apparently the references seem to be indeed a better indicator for topical relatedness than citations. Their application for exactly that purpose in this thesis is explained in more detail in Section 5.2.

3.3.4 Journal to Journal Citations

As a further option for a disciplinary organization of science, journal to journal citations can be used. Then, the topical clusters are pre-defined by a journal's context. In this way, development trends and convergence of topics can be studied (Doreian and Fararo 1985, Borgman and Rice 1992, van den Besselaar and Leydesdorff 1996, Leydesdorff 2003).

Klavans and Boyack (2006a) evaluated ten different variations of journal relatedness measures based on citations. Six of these measures were based on journal to journal citations, the remaining four use co-citation of journals as a similarity measure. In their study, journal to journal citation based measures outperformed the co-citation metrics, but they acknowledged the usefulness of co-citation based measures when information about journals is needed for which no citation information but only cited information is given.

Archambault, Beauchesne and Caruso (2011) listed various limitations to journal-based classifications and therefore developed a taxonomy (called "ontology" by the authors) to which the individual journals were assigned.

3.3.5 Author Networks/Communities

There are various possibilities to investigate the development of author networks. First, community detection systems and other social network analysis methods can be used to investigate the overall structure of an author network. Such a network can be constructed with graphs in which the nodes represent the authors and the edges are co-publications. The thickness of the edges can represent the number of co-publications a pair of authors has written in a specific time span. The second option is a qualitative analysis of an author network to derive incentives and reasons for collaborations in this network. In this subsection, a brief overview of both options is given.

A quantitative, graph-based analysis of an author network was performed by Newman (2001). The graphical representation allows the adoption of graph-based metrics. Examples for these metrics used are "transitivity", "betweenness" and "collaboration weight". The transitivity measures the degree to which a connection between authors a_1 and a_2 and a_2 and a_3 implies a connection between a_1 and a_3. The betweenness is the number of shortest paths between any pair of authors a_1 and a_3 that passes through author a_2. Thus, this metric reflects the number of authors that are connected via a third author without whom they would have no connection at all. The "collaboration weight" refers to the weighted number of collaborations of two authors a_1 and a_2 that shows the extent of their collaboration. Despite his findings, Newman (2001) also referred to some aspects of the graph model that demand a social explanation.

Racherla and Hu (2010) conducted a similar study restricted to the field of tourism research. With the help of similar metrics as Newman (2001), they were able to deduce the meaning of collaboration in this specific research field. Even though some of the most prominent researchers did not show a high betweenness value, the establishing of a big collaboration network may correlate with a high productivity. Again, the limitations of a solely quantitative analysis become evident, since it could not be deduced whether a high productivity causes other researchers to collaborate more often or whether the collaborations enable the high productivity.

Of course, such a network analysis is not restricted to the author level. In other studies, the collaboration pattern analysis was conducted on the level of NUTS[52]-regions (Hoekman, Frenken and Oort 2009), countries (Li et al. 2010) or fields (Rosvall and Bergstrom 2008).

Pepe and Rodriguez (2010) analyzed the development of one single research institute to look for changes in the collaboration behaviour. With their qualitative analysis it is possible to interpret changes like e.g. an increase in interdisciplinarity of the research performed.

An interesting approach to detect hidden collaboration potential was proposed by Giuliani, Petris and Nico (2010). They measured collaborations by co-authorship and content of publications by keywords used in these publications. They were able to reveal unused collaboration potential for authors in a research group. They detected these potential collaborations by identifying authors that shared the same keywords in their publications, thus apparently worked on a similar topic. In this context, they were also able to confirm the role of geographical distance in collaboration. Comparatively, Hoekman, Frenken and Oort (2009) showed that even though "the choice for a collaboration partners [sic] should be based solely on scholarly ground, [...] this choice is significantly impeded by geographical barriers" in the case of the EU (Hoekman, Frenken and Oort 2009, p. 736). Similar results have been found for collaboration of firms (Torre 2008).

Ding (2011) extended an LDA approach to detect communities in a co-authorship network. Basically, this is the reversal of the approach explained in this thesis (see Chapter 5), where an LDA model that is extended by the author distribution, detects clusters of documents. Yan et al. (2012) in turn extended the approach by Ding (2011) with the inclusion of the publication source. Thus, the LDA model does not only use the textual information and author names but also the "stamp" of a journal or conference. With their approach, they can show for the case of Information Retrieval that "topics are sustained by the creation of a community around these topics" (Yan et al. 2012, p. 152).

3.3.6 Keyword Networks, Text-Based Clustering and Co-word Analysis

The similarity between text documents can be measured by the number of words/terms they have in common (see e.g. van den Besselaar and Heimeriks 2006). The usual procedure is to calculate the term vectors of the documents and then compare these vectors according to a similarity measure. Börner,

---

52 Nomenclature of Units for Territorial Statistics: "[A] hierarchical system for dividing up the economic territory of the EU", http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction, last accessed on 2014/05/23.

Chen and Boyak (2003) introduce the metaphor of comparing DNA fingerprints when talking about key term comparison, since a higher similarity between them corresponds to a higher probability of both being associated with the same "species".

The similarity between two term vectors can be calculated by different measures, the most common being the Cosine Similarity and the Jaccard Index (cf. Tan, Steinbach and Kumar 2006).

According to the work by van Eck and Waltman (2009), the most appropriate metric to measure term co-occurrence is the equivalence index[53] (cf. Cobo et al. 2011, p. 149). Opposed to the aforementioned similarity metrics between term vectors, this metric measures the similarity between two keywords w_1 and w_2. The more of the documents that contain keyword w_1 (M_1) also contain keyword w_2 ($M_{12} = M_1 \cap M_2$), the higher the similarity between both keywords.

$$e_{12} = \frac{M_{12}^2}{M_1 * M_2} \text{ (6, equivalence index)}$$

Thus, two kinds of co-word analyses can be differentiated: 1) a document based approach in which the similarity between documents is calculated based on shared words and 2) a keyword based approach where the similarity between keywords is calculated based on the number of shared documents (normalized with the total number of documents in which they appear).

The former was for instance applied by Boyack et al. (2011) to compare different keyword-based approaches for clustering MEDLINE articles. According to Cobo et al. (2011), co-word analysis results in clusters of scientific publications representing the main concepts of a field, while co-citation analysis shows the overall structure or connections within a field. The distinction between the two results can be illustrated by the difference between the analysis of cities or urban agglomerations and that of the routes between them. The focus of co-word analysis on topic representation justifies the high dependency on term usage in this thesis to build topic clusters.

The other form of co-word-analysis is the clustering of terms according to the number of shared documents. Thus, terms that oftentimes appear in the same documents are grouped together. This kind of co-word analysis was primarily introduced by Callon et al. (1983). It was adapted to the analysis of various topics (see Whittaker 1989, He 1999, Cahlik 2000, Coulter, Monarch and Konda 1998, to name only a few). Yi and Choi (2012) and Courtial (1994) calculated keyword networks for publications in selected journals and were thus able to look for trends and developments in these journals. Coulter, Monarch and Konda (1998) and Courtial and Michelet (1990) showed that this kind of co-word analysis is in general useful for monitoring the development of a scientific discipline over time to account for changes in the vocabulary or a shifting focus. The resulting keyword network can be analyzed by the usual network indicators like density and centrality to assess individual keyword clusters (Callon, Courtial and Laville 1991, Muñoz-Leiva et al. 2012). On a similar notion, the usage of terms during the emergence of topics in comparison to the remainder of the dataset and later years is

53 Also referred to as association strength, proximity index and probabilistic infinity index.

analyzed in Chapter 10 to confirm the assumptions on term usage underlying the LDA approach (see Chapter 5).

3.3.7 Research Fronts

Price (1965) defined the research front in science as a set of papers "knitted together by the new year's crop of papers" – the "epidermal layer" of science (Price 1965, p. 512). Given this definition, a research front can easily be tracked by the references in a field, where research front papers are referred to by half of the current papers in a field (Price 1965). The delineation between a "hot topic" and a research front is fuzzy, as a communication between Leydesdorff and others made clear.[54] Both refer to topics on which the majority of scientists in a field are currently working on.

Whereas Price only used direct citations in his initial study, there have been other studies that also used or compared bibliographic coupling and co-citation for this task (Kajikawa and Takeda 2009, Shibata et al. 2009a). Furthermore, other indicators like the authors' publication output (Tsai 2011) or the states in a Marcov model (Lee, Lee and Yoon 2011) have been used to judge the development stage of a topic.

As stated by Kajikawa and Takeda (2009) the use of co-citation results in a time-lag between the publication date of the clustered documents and the point in time where an analysis is possible. The usage of direct citations on the other hand demands an already well-connected topic for which enough citable, relevant literature is available and detectable. An alternative is the use of bibliographic coupling, which uses only the references of the documents as features for the clustering. However, this kind of usage is a clear contradiction to the initial definition in which the citedness of a document also is an indicator of its innovation factor. Thus, this aspect of the analysis is lost and it is reduced to a mere clustering of documents for which other indicators have to be used to decide whether a document cluster represents a research front or not. In this thesis, the usage of citations is excluded because of the already mentioned induced time lag. Nonetheless, the aim and thus also the results of both analyses are comparable.

3.4 Summary

This chapter presented various aspects of bibliometric analysis and its relationship to this thesis. The databases used in this thesis were illustrated in more details. Section 3.2 discussed popular indicators in bibliometrics in general. Again, a closer look at citation analysis was necessary to distinguish between the usage of references – as in this thesis – and citations for the calculation of topical relatedness. The JIF was introduced as one indicator for journals as it is also used in the later development of rules for the detection of emerging topics in this thesis (Chapter 6 and Section 9.4). A list of possible metrics for the mapping of scientific publications in general showed the alternatives as well as parallels to the approach presented in this thesis.

---

54 http://listserv.utk.edu/cgi-bin/wa?A2=ind1010&L=SIGMETRICS&D=0&m=5305&P=15045, last accessed on 2012/11/05.

4 Machine Learning Foundations

This chapter gives an overview of the concepts and terminology in Machine Learning that are of use for the remainder of this thesis. The chapter focuses in particular on the distinction between Training and Test Sets, metrics for the automatic qualitative assessment of Machine Learning approaches and similarity calculations. In the final Section 4.3, a bridge is built between the theoretic foundations and their implementations in this thesis. The respective transfer concerns only the similarity calculation, but the following Chapter 5 covers the other main part of the approach, namely LDA.

4.1 Terminology

4.1.1 General Definition and Example

Machine Learning denotes the progress of a computer program improving its performance for a specific class of tasks with experience (Mitchell 1997, p. 2). This improvement is oftentimes achieved by deriving rules from different occurrences of observations. These rules should, on the one hand, cover most of the observations in the dataset but, on the other hand, hold in general settings. In this way, the performance in regard of the tasks improves with the level of experience (Mitchell 1997, p. 2).

A Machine Learning approach defines how rules for a specific task are derived. The approach is independent of already existing rules or the task at hand. Thus, the specialization to a specific task is a subsequent step. In this way, Machine Learning algorithms can be adapted to various classification or distinction problems. In other words, there is one stable part or concept of the algorithm that denotes how the learning takes place and there is one variable part that is concerned with the adaptation to the specific problem. Turing (1950) described the distinction between these two concepts as follows:

“The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States” (Turing 1950, p. 458).

As one example, Neural Networks are learning algorithms that calculate result values for a set of input values. These calculations are based on connections between neurons that are established during the learning phase.

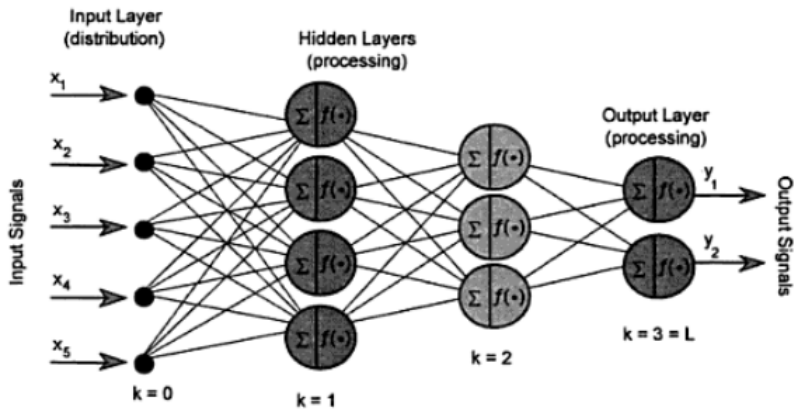
The concept of a Neural Network and its applications will serve as a running example in this section to illustrate the concept of Machine Learning. Basically, Neural Networks can be applied for all problems for which a decision has to be derived based on a multiple set of (inter-dependent or independent) factors, for example, decisions for steering a vehicle (Kehtarnavaz and Sohn 1991) or the pronunciation of a letter in an English text (NETtalk by Sejnowski and Rosenberg 1986, also described in a

demo for neural networks[55]). For the latter example, the input for the neurons in the input layer are the respective letter itself plus the six surrounding letters, i.e. the three preceding and the three following letters, which are represented by 203 units (26 letters plus 3 units to encode punctuation and word boundaries times 7 positions). By combining these input features, the phoneme is determined, which in turn is represented by a combination of 26 units in the output layer.

The concept of Neural Networks is based on biological neurons, which "fire" as soon as a certain threshold of necessary transmitters is achieved. The same principle can be used for Artificial Neural Networks, where each neuron gets input values from different other neurons or the input layer (comparable to human perception or reaction that results in transmitter emission for the neurons). Thus, the total input of a single neuron is the sum of these various sources, where a different weight might be assigned to each of them. A neuron "fires" as soon as the threshold of the neuron is reached: it sends new transmitters along to the next layer. Note that the amount transmitted by the neuron, the "firing", is not the amount that was taken in nor is it the difference between the input and the threshold. It is a binary value that is either 0 (if the threshold was underscored, i.e. no "firing") or 1 (if the threshold was reached). However, this output might be weighted in turn before it is fed as an input to a subsequent layer's perceptron. The weight between two connected neurons might also be 0 if actually no interaction between these two neurons is sought.

In a Neural Network, the number of hidden layers and units are provided by the user, but the weightings between the units are determined during the training of the Machine Learning approach. Figure 16 shows one possible instantiation of such a network. This network has one input, two hidden and one output layer. In general, only the number of input and output layers is restricted to one. The numbers of hidden layers as well as the number of neurons in each layer are completely flexible.

Figure 16: Exemplary Neural Network.



Source: Priddy and Keller (2005, p. 8)

The next subsections briefly describe the various training methods in Machine Learning and illustrate them in terms of the running example, the pronunciation of letters.

55 http://ecee.colorado.edu/~ecen4831/lectures/NNdemo.html, last accessed on 2013/04/28.

4.1.2 Instances, Training and Test Set

In Machine Learning, the subjects of an analysis are denoted as “instances”. In the case of this thesis, they represent scientific publications. Each instance can have a number of features that are included in the analysis. If the Machine Learning task is a classification problem, one of these features can be the class value. Then, the goal is to learn the relationship between the class value and the other feature values so that the class value can be derived for new instances for which it is (yet) unknown. In this thesis, Machine Learning is concerned with the relationship between the bibliometric features of a document and the development status of its topic. Thus, the algorithm determines whether a document belongs to a new topic or not.

To apply and test the Machine Learning approach, the so called “Training Set” and “Test Set” are needed. Based on the Training Set, the Machine Learning approach derives rules between the features of the instances (in the example above, the surrounding, position and combination of the letters) and the correct solution (the phoneme). In the pronunciation example, the goal value is a kind of labelling or classification (e.g. “hard c”). In other cases, called regression problems, the goal value can be numeric. The number of features F determines the dimension of a vector that is mapped to a value y. All characteristics of an instance that are not covered by the F-dimensional feature vector and the goal value y are ignored by the Machine Learning approach.

The acquisition of a good Training Set is the basis of a well-performing Machine Learning approach. All rules that are not covered by examples in the Training Set, are not learned and thus not applied if a respective case is encountered later. In the running example, the algorithm is supposed to learn that ‘e’, ‘i’ or ‘y’ following a ‘c’ or ‘g’ means a soft pronunciation of the latter, a hard pronunciation otherwise.[56] Thus, at least one word for each combination of these letters (and one for all the remaining variations) should be included in the Training Set. Otherwise, the pronunciation of the letter will be determined based on the remaining information and therefore may be wrong.

Thus, the Training Set serves to tune the approach. In order to test the resulting approach, a second dataset is needed. Otherwise, the approach could perform especially well on the Training Set as it learned Training Set specific instead of general rules. The most extreme case would be a set of rules which specifically maps the input values of each instance in the Training Set to its output value. The phenomenon that a Machine Learning approach works better on the training data than on general data is called “overfitting” (Witten and Frank 2005, p. 86). This concept underlines the importance of a sufficient Training Set selection.

As explained above, rules, which are not represented by training examples, are not derived. Similarly, if the Training Set contains misleading examples (e.g. a “c” at the beginning of a word is always pronounced “hard”), generally non-applicable rules are generated. The evaluation on the Test Set checks whether such misconceptions are represented in the rules. Therefore, the now static approach is run after training on the Test Set. Thus, it is no longer learning but its performance at the current state is assessed. For this purpose, the Test Set usually also contains data for which the performance can be

56 http://esl.about.com/od/speakingintermediate/a/hardsoftcg.htm, last accessed on 2013/04/28.

easily measured, i.e. instances for which either already a correct solution is known or for which the results gained with the Machine Learning approach can be manually assessed.

There are different approaches how to "learn" with a given Training Set. Supervised learning assumes that the correct solutions for a given task are (at least) partially known, i.e. there is already a mapping of the input values $x_1,..., x_n$ to a possible output value y for at least some instances. The goal is to derive a set of rules that applies generally and is thus not dependent on the explicit expression in already observed cases.

The Training Set contains those examples with already observed values for y. This might be for instance a set of plant characteristics for which the fitting plant label is already known or customer data for which the insurance rate was calculated before. Or, to stick with the running example, a set of words or texts for which the pronunciation of the letters is represented in the desired form.

Alternatives are unsupervised learning programs, in which no goal value exists for the instances in the Training Set. The most common examples are clustering programs. It might be already known that letters and their features should be classified according to their pronunciation (e.g. 'cAt' and 'lAck' together) in a certain number of groups, but the "how" is left to the Machine Learning approach. A clustering approach uses the features of a set of instances to calculate the similarity between these instances and groups them so that the instances with the highest similarity end up in one group, a so called cluster.

There are variants of Neural Networks which use supervised learning, e.g. the Self-Organizing Maps (Kohonen Maps, Kohonen 1982). For these, the weightings, which are initially assigned randomly, are adapted by repeated processing of the training data. In this way, the Neural Network by and by adjusts to the underlying training data.

The distinction between a clustering approach and a classification is that for the latter, the groups of instances for one topic are not only aggregated but also labelled in respect to the specific class. However, both aim at an aggregation of similar instances or, in the case of this thesis, documents. In the classification problem, the features of the documents that fall under one label are fixed. Thus, it is already known which kinds of "buckets" exist and for a small set of instances (the Training Set) the respective buckets are known. The goal of the Machine Learning approach is thus to derive, how the instances are connected to the respective class. In an unsupervised clustering problem, neither the "buckets" nor the method of assigning the instances to these buckets are known. Rather, this issue is solved by the Machine Learning algorithm. Nonetheless, both learning methods are comparable and it will be explained later how these common characteristics can be used to measure the performance of the clustering in this thesis.

For the sake of completeness, reinforcement learning, the middle ground of both so far mentioned learning methods, is briefly explained as well: In reinforcement learning – as in unsupervised learning – the training data does not contain the goal values. However, the result generated by the approach in each iteration is assessed on the fly and delivers a cost or reward value. In this way, the learning algorithm can approach the correct solution by and by. Thus, "the learner is not told which actions to take, as in most forms of Machine Learning, but instead must discover which actions yield the most reward

by trying them", which results in a "trial-and-error search" for the right behaviour (Sutton and Barto 1998, p. 3).

4.1.3 Evaluation Metrics

For the evaluation of a Machine Learning approach, Precision, Recall and F-Measure can be calculated. All three metrics stem from Information Retrieval and thus make only a distinction between documents that are sought by the end user (positive instances) and other documents (negative instances). Usually, the basis of such an evaluation is the result set, i.e. the set of documents that are returned to the end user with a certain system and/or query. However, the comparison of the metrics for the result and the initial dataset (if possible) show whether the respective system is indeed an improvement to the status quo.

Recall denotes the ratio of the number of documents an Information Retrieval system correctly identified to the maximum number of documents it could have (correctly) found. Thus, it reflects the share of positive instances that are returned to the end user in relation to the true number of positive instances (Baeza-Yates and Ribeiro-Neto 1999, p. 155). Or in other words, the Recall is "an estimate of the conditional probability that an item will be retrieved given that it is relevant" (van Rijsbergen 1979, p. 120). It follows, that the positive instances must be known in advance for its computation. The respective Training Set should thus cover this information. This is usually achieved by a binary value of *y*, that indicates if the instance is positive or negative. If the best solution for a retrieval problem is known, the respective dataset is called Gold Standard. The overlap between the document set found by the Information Retrieval system and the Gold Standard set can be counted. These documents are labelled "true positive" since they were found (positive) and were supposed to be found (true). In contrast to that, the documents that are included in the Gold Standard but not in the result set are called false negative. Those documents that were returned by the system even though they do not fulfil the right criteria (according to the Gold Standard) are labelled "false positives". The false positive documents impurify the result set because they would make it harder for the end user to find the sought-after documents in the result set.

The Recall can then be calculated by dividing the number of positive instances in the result set by the total number of positive instances. In total, the Recall denotes the share of documents from the Gold Standard that were found, which can be expressed by (Baeza-Yates and Ribeiro-Neto 1999, p. 155):

$$R = \frac{M_{pr}}{M_{pi}} \quad (7)$$

where *M* denotes the number of instances. The annotation *pr* indicates that only the positive instances in the result set of the system are examined, while *pi* refers to the positive instances in the initial dataset. Thus, the ratio of the positive instances in the result set ("true positive") and in the initial set are compared.

On the other hand, the Precision of a retrieval system sets the number of true positive documents to the total number of documents in the result set (Baeza-Yates and Ribeiro-Neto 1999, p. 155):

$$P = \frac{M_{pr}}{M_{pr}+M_{nr}} \quad (8)$$

Again, M_{pr} denotes the positive instances in the result set, while M_{nr} are the negative instances that are returned to the end user.

Thus, the Precision of a dataset denotes its purity, i.e. the share of positive documents in the result set (Baeza-Yates and Ribeiro-Neto 1999, p. 155). The Precision can be also seen as "an estimate of the conditional probability that an item will be relevant given that it is retrieved" (van Rijsbergen 1979, p. 120). Thus, for both Recall and Precision, the documents that are relevant for the end user need to be already known.

If the system simply returns all available documents, the Precision will be equal to that of the initial set. For only positive instances the Precision is 1, for only negative instances it is 0. If there are 100 documents of which only one is in the Gold Standard, the Precision would correspond to 0.01. This accounts for the fact that the end user would have to read 99 documents – in the worst case – before finding the correct one. To take into account rankings, sometimes Recall @ *n* and Precision @ *n* are calculated which represent these metrics calculated for the top *n* ranks in the result set. Thus, Precision @ 1 is either 0 or 1 depending on the fact whether the/a right document was returned at that position or not (see Witten and Frank 2005, p. 171).

The F-Measure is the combination of both metrics, so that a fair balance between coverage of positive instances (Recall) under the exclusion of negative instances (Precision) can be found. The formula for the F-Measure is (cf. van Rijsbergen 1979):

$$F_\beta = \frac{(1+\beta^2)(P*R)}{(\beta^2*P+R)} \quad (9)$$

With β being a positive, real number, *R* being the Recall and *P* being the Precision (Witten and Frank 2005, p. 172).

Using the F-Measure as a basis of the evaluation and selection of the retrieval system avoids extreme rules, where the result set might be

- small, but very pure, leading to a high Precision but a low Recall, or
- large, covering many or all positive documents, but also many negative ones, accounting for a low Precision and a high Recall.

In contrast to that, result sets with both a high Recall and Precision should be preferred and this is why the F-Measure should be applied. In this thesis, the $F_{0.5}$-Measure is used, which weights the Precision higher than the Recall. Thus, approaches with a pure(r) document set get a higher score even though they might also have a smaller coverage. They are preferred to sets with a high Recall but also more negative instances (old documents). The intention of the overall approach is a method that delivers an as pure as possible set of new documents to the user. If a full coverage is desired, the initial dataset

could be used. By using β = 0.5, half "as much importance to Recall as Precision" (van Rijsbergen 1979, p. 133) is given.

Therefore, the quality of the representation of topics and the purity in the clustering should be measured. Both measures of Recall and Precision were adapted for the evaluation of clusters in this thesis in accordance with the definitions of "purity" and "Entropy" by Baeza-Yates and Ribeiro-Neto (1999, pp. 357ff). The Recall of a topic is used to calculate how many documents of a topic are covered by a single cluster. The aim is to represent each topic in exactly one cluster. Thus, for each topic *t*, the cluster *k* that contains most of its documents is selected for the evaluation. The share of documents of topic *t*, which is covered by *k*, represents the Recall of this cluster *k* for topic *t*. To evaluate the whole result set, the average Recall over all topics is calculated as

$$R_K = Avg_T \frac{max_k (M_k \cap M_t)}{M_t} \quad (10)$$

R_K is the adapted Recall for the clusters *K*, *t* is a topic (in the set of topics *T*) defined in the Gold Standard and *k* is a cluster (in the result set *K*). If each topic *t* is represented by exactly one cluster *k*, the Recall corresponds to 1.

There is no weighting of the topics according to their size as this would imply that the importance of the correct representation of a topic increases with its size. That would be counterproductive for the system as it should also (and especially) represent topics that include only few documents. Thus a weighting according to size in terms of documents is not applied. However, a bigger topic is more difficult to cover completely in one single cluster than a smaller one as there are more aspects and documents to include per se.

With the foregoing definition of Recall, it is measured how close the clustering result comes to representing each topic by a single cluster. However, in the ideal case each cluster should also represent only one topic. Thus, the Precision measures the share of documents in a cluster belonging to its dominant topic, i.e. the topic of which it contains most documents. Again, for the overall evaluation the average is used – this time across all clusters:

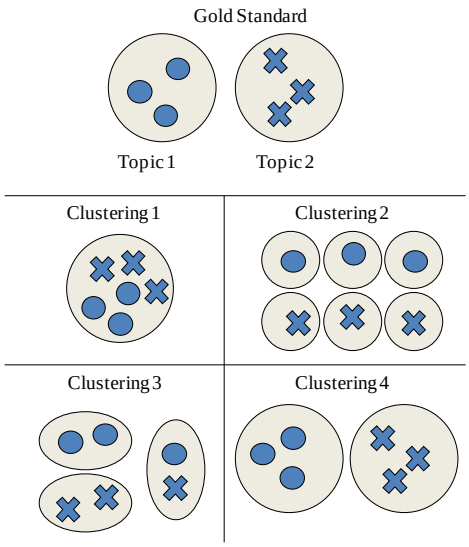
$$P_K = Avg_K \frac{max_t (M_k \cap M_t)}{M_k} \quad (11)$$

P_K is the Precision for a set of clusters *K* and again *k* and *t* represent a cluster or topic respectively. The measure calculates how "pure" a cluster is and corresponds to the metric "purity" given by Baeza-Yates and Ribeiro-Neto (1999, p. 357).[57] Thus, a cluster that contains only documents from one topic results in the Precision value 1, a cluster *k* with M_k documents where each document stems from a different topic equals the value $1/M_k$. Figure 17 gives an example of possible clusterings of a set of six documents and the comparison with the Gold Standard in terms of Recall, Precision and F-Measure (for β = 1 and β = 0.5).

---

57 Given how the two metrics Purity and Precision calculate the same metric but only on different aggregation levels, it seems appropriate to use only one notation for both in this thesis.

Figure 17: Examples for possible clustering of six documents and the resulting values for Recall, Precision and F-Measure.



Clustering	Recall	Precision	F-Measure	$F_{0.5}$-Measure
1	1	1/2	¾	5/9
2	1/3	1	2/3	5/7
3	2/3	5/6	20/27	25/36
4	1	1	1	1

Source: Own illustration

Another metric in Machine Learning is the *Information Gain*, which is calculated as follows (see Moore 2003)

$$IG(class|feature) = H(class) - H(class|feature) \quad (12)$$

where the *class* is one of the possible classifications in the dataset (e.g. in this case "new"), the *feature* is one of the features of the instances (e.g. the size of the journal) and *H* is the "Information Entropy" either in a class, *H(class)*, or in a class if the feature value is known, *H(class| feature)*. In turn, for a *class* with possible values *J* the Entropy is denoted as

$$H(class) = -\sum_{j \in J} p_j log_2 p_j \quad (13)$$

Or with regard to a *feature* with possible values *X*

$$H(class|feature) = \sum_{x \in X} p_x H(class|feature = x) \quad (14)$$

In the above formulas, note that the probability for a class values p_j and that for a feature value p_x are calculated for the whole dataset, while the Entropy $H(class|feature = x)$ uses Formula (13) to calculate the Entropy for the class in the subset of all instances with the feature value *x*.

Informally speaking, the Information Gain denotes how many bits (as a measure of amount of information) in a communication can be omitted if the feature value is known by the recipient.[58] While the Entropy of a class, *H(class)*, shows how much information is needed for communicating the class according to the initial distribution, *H(class| feature)* denotes the amount needed when the feature value is already known. For example, if the initial distribution of the classes new and old was uniform,

58 See Moore (2003).

1 bit would be needed to communicated the class value of a document (1 for new, 0 for old or vice versa): $H(class) = 0.5$. However, if the specific feature value of an instance bears information of its class, the direct communication or the transfer of the bit regarding the class might be obsolete. For instance, if it was known that for values of journal size smaller than 500 documents per year the class is old, no (further) information needs to be communicated. For the simple case that all other documents are of the class new it holds that $H(class|journalSize) = 0$ (the Entropy is zero because the class value is already observable) and $IG(class|journalSize) = 0.5 - 0 = 0.5$. Thus, with solely the knowledge of the feature value, all necessary information can be derived and the communication is reduced to 0 bits.

However, in order to apply the Information Gain in this thesis, the distributions of new and old documents had to be smoothed. Otherwise, evaluation metrics like the Entropy (and also Precision and Recall) are influenced by the overall distribution. If the a priori distribution is skewed, i.e. there are much more instances for one class than for the other, the Information Gain will always be low as certain values are more probable and the overall Entropy is low.

For a low Entropy, *H(class)*, the leeway for improvements through the usage of additional attributes, is relatively small. In general, the Information Gain can be at maximum $IG = H(class)$. The minimum value is $IG = 0$, which corresponds to a known class value for an observation like in the previous example. Thus, the initial value *H(class)* also determines the possible value of an attribute's usefulness (i.e. the Information Gain); for an initially small probability of a class, the Information Gain will always be very low. That, however, would be counterproductive for this study as these a priori probabilities are irrelevant for deciding whether a document is an innovative paper. In theory there should be a 50% chance that the approach labels a document as new. In practice, the chance that a labelling as new is false corresponds to the initial share of new documents in the dataset. A dataset skewed in favour for old documents would lead to a system which is more prone to classify a (new) topic as an old one, since the chance for failure in this case is always smaller than for a classification as new. This would lead to more false negative classified instances, but due to the small portion of new instances in the dataset also to high evaluation metrics.

Table 5: Possible classifications of topics in the use case.

Class	**Classified as**	
	New	**Old**
New	True positive, $p = \frac{M_n}{M_o+M_n}$	False negative, $p = \frac{M_n}{M_o+M_n}$
Old	False positive, $p = \frac{M_o}{M_o+M_n}$	True negative, $p = \frac{M_o}{M_o+M_n}$

Source: Own illustration

Table 5 shows the two classes (new and old) which a topic can have and be allocated to. In a set of M instances, the probability of an instance being new is $p = \frac{M_n}{M} = \frac{M_n}{M_n+M_o}$ where M_n or M_o is the number of new or old instances. For a diminishing share of new instances M_n in the dataset, the probability for true positive assignments decreases. A classifier would incorporate this by classifying as old when "in doubt". Also, for a small M_n, the evaluation metrics are good even if the classifier labels all instances

as old. The percentage of correct classified instances would then be at $\frac{M-M_n}{M} = \frac{M_o}{M_o+M_n}$ which approaches 1 when M_n tends to 0. In this way, a skewed distribution leads to a biased classifier. Thus it is more reasonable to use a smoothed distribution. By doing that, attributes for new documents are equally considered in the classification process.

A "smoothed distribution" can be achieved by Resampling, which is used to prepare the dataset on which the rule deduction in Chapter 9 is applied. For Resampling, a dataset is generated by drawing randomly with replacement from the original dataset. The size of the new dataset can be parameterized. In the case of this thesis, it was always as large as the initial dataset. However, the main point of this method is that the drawing can be "manipulated" to change a biased class distribution to a uniform one. In this case, a drawn instance is put back without adding it to the new dataset if it contradicts the goal of a uniform distribution. Thus, drawing for example coincidently repeatedly from only one class is unproblematic as long as a uniform distribution can still be achieved, but after that, all instances drawn from the respective class are omitted. Note that a uniform distribution does not necessarily demand an exact partitioning of the dataset according to the number of classes but allows for some variance. Furthermore, a dataset size equal to that of the initial set requires that one if not more instances of the underrepresented class are included multiple times in the new set. Because of the replacement in the drawing procedure, in theory each class might be represented by only one instance that is multiplied as often as the dataset size demands it.

4.2 Similarity Measures

Clustering is the process in which similar instances are grouped together to form one set, a so called cluster. In order to measure the similarity between instances in a formal way, a similarity function is applied. This similarity function can determine the similarity of pairs of instances and/or clusters.[59] The (groups of) instances with the highest similarity at a time are then grouped together to form one single cluster. In theory, the initial clustering between instances is a clustering of clusters, which have a size of 1 each. For each cluster pair, the similarity is calculated and the clusters with the highest similarity are merged. Then, the process is repeated until the stop criterion is fulfilled. The fact that the size of the single clusters increases with time is irrelevant as the similarity function should consider both single-instance as well as multi-instance clusters.

After the clusters in different time periods have been built, the system in this thesis identifies topically related clusters. The method for this corresponds to that used in the clustering, as the clusters are compared based on features to calculate a similarity value. However, in this case the clusters with a high similarity are not merged but linked across annual boundaries.

The similarity between instances can be calculated by different combinations of the respective features. For this, the *F* features of each instance can be represented as vectors in an *F*-dimensional space.

---

59 This paragraph only deals with hierarchical agglomerative clustering. Agglomerative clustering starts with a cluster for each instance. With each iteration the clusters are merged. In contrast to that, other clustering approaches start with one single cluster containing all instances which is iteratively split in smaller clusters.

The easiest way is to aggregate over the similarity comparison of the individual feature pairs. For instance, the sum over the absolute difference in feature values *x* for the features *F* can be calculated:

$$Sim(d_1, d_2) = \sum_F \left| x_{1f} - x_{2f} \right| \quad (15)$$

As an alternative, the Euclidian Distance between the two vectors can be used:

$$Sim(d_1, d_2) = \sqrt{\sum_F (x_{1f} - x_{2f})^2} \quad (16)$$

This procedure might also use weightings of the different features. For example, the adaptation of the above function to

$$Sim(d_1, d_2) = \sum_F w_f * \left| d_{1f} - d_{2f} \right| \quad (17)$$

Furthermore, individual distance/similarity functions for each feature can be introduced to account for differences in the features:

$$Sim(d_1, d_2) = \sum_F w_f * sim_f(d_{1f}, d_{2f}) \quad (18)$$

This can be especially useful when mixing nominal and numeric features, e.g. a document's statistic and its terms. An exemplary vector might look like this:

$$\begin{pmatrix} length\ in\ words \\ age \\ number\ of\ accesses/citations \\ title \\ authors \end{pmatrix}$$

The similarity in the titles might in turn again use a vector representation of the terms, where each position of the vector denotes the usage of a certain term. Thus, the actual information about the order of the terms is lost. Such a vector representation of a set of terms, i.e. $(w_1, w_2, w_3, \ldots)^{-1}$, is called "bag of words".

The vector representation of the features or terms makes it possible to calculate the Cosine similarity between them. Thus, they are indeed used as vectors in an *F*-dimensional space between which the degree is measured.

In this thesis, the similarity between the instances and clusters during the clustering is calculated in a topic model, which uses the textual information as a bag of word. The underlying calculations, which enable the clustering of the instances, are explained in Chapter 5. Thereafter the connections between the clusters in different time periods are established. The necessary similarity calculation uses the term distributions of the LDA approach and is explained in full detail in the next section.

4.3 Similarity Calculation of Topics in this Thesis

As stated above, the resulting clusters are compared across the different time periods. The preceding clustering is performed for each year separately to allow for trends to be discovered. If a longer time

period was used, the weak signals for trends might be eclipsed with those of more established topics. Thus, the dataset is split in yearly portions and then the topics in these parts across the yearly boundaries are compared. The similarity between the clusters can be measured with the aggregated term distributions, references, authors involved in the topic etc.

Small (2006) used a similar approach but applied overlapping time windows to extract the connections between document clusters in different years. Thus, he calculated the clusters in a period between year *y* and year *y+2* to look for documents that were connected over these years. As stated above, this is not an option when topics are yet emerging and the observation period ends with the year for which the emerging topics are to be detected. If a topic is emerging, there should be no connection to any previous topics (or only weak ones). But if only one document represents a new topic, the chance of "burying" this document in another topic increases with the inclusion of former years. This effect is enforced as the vocabulary used in the beginning of a topic might still resemble in large parts that in other topics.

When two clusters from different time periods are compared, the necessary delimitations of the topics have already been made by the clustering. Now the relations on the more aggregated topic level are detected in the similarity calculation.

The new dimension time enables some additional features that were impossible before: On the one hand, references to documents in other clusters provide a clear hint on relatedness.[60] On the other hand, the time elapsed between two topics can be used as a kind of discount factor for their relatedness or the probability that the later cluster is directly founded on the former one.

The similarity between the clusters can thus be determined based on the following features:

- common terms/vocabulary comparison
- common references
- distance in years

Co-citations and similar metrics are not available for the documents in the most recent, i.e. current time period (see in particular Section 3.2.1), and thus not covered in the following.

For the first two features, the similarity calculation uses the term and reference distributions of two topics; the term and reference vectors of both clusters are compared.[61] Basically, the term vector is a *V*-dimensional vector for *V* different unique terms in the whole document set. Element *w* of the vector denotes the absolute frequency of a term. The term vector for a document shows the absolute term frequency of a term *w* in respect to a specific document, i.e. restricted to occurrences in this document. The term vector of a cluster denotes the number of times the *V* terms appear in all the documents of that cluster. Two term vectors of two clusters can be compared by the cosine similarity, i.e. assuming that both vectors are depicted in a *V*-dimensional space, the degree between them is calculated. The

---

60 Contrary, the absence of citations does not necessarily correspond to topical distinction but can have various reasons, see Section 3.2.1.

61 More precisely, the calculation is based on the distributions T and φ of LDA, which is explained in Chapter 5.

same can be done for the references. The combination of both metrics is achieved by a weighting w_r that denotes the relation between both values. The similarity in respect to terms cannot be used as the sole base of cluster comparison because the vocabulary of a topic changes over time, new specific terms are introduced and older terms and the respective concepts are abolished. Thus, the overall similarity between two topics t_1 and t_2 can be described as

$$sim(t_1, t_2) = w_r * sim_r(t_1, t_2) + (1 - w_r) * sim_t(t_1, t_2) \quad (19)$$

where $sim_r(t_1, t_2)$ and $sim_t(t_1, t_2)$ is the cosine similarity between the two reference and term vectors and w_r is a value in [0;1]. If the same vocabulary and references are used in both clusters, then it can be assumed that both clusters are topically related. The similarity value as a value between 0 and 1 denotes the degree of agreement between the two sets of vocabulary and references.

The distance in years refers to the time span between the two compared topics. The reasoning behind this feature is that the chance of a topic reappearing after a number of years *Y* (without interim linking topics) should diminish over time and thus with increasing *Y*. Since the approach is intended for detecting new emerging topics in the most recent year, the distance in years between the clusters which are compared corresponds to the age of the older topic.

After the similarity between clusters in different years has been calculated, links are built between those clusters with the highest similarity or a similarity higher than t_c. These links represent a topic evolution, i.e. the connection of two clusters in different years indicates that the former topic evolved to the newer one. Thus, by linking those clusters with a high similarity, their development over time can be monitored. Also, this approach makes it possible to detect those topics that have no predecessor in the sense of this linkage. When considering the way these links are established, this means that the topic did not occur from “thin air” but that all other topics (those from which it evolved as well as those, which are completely unrelated) do not share a similarity that is above the threshold t_c and are thus not sufficiently similar.

An alternative or in the case of this thesis complementary approach would be the detection of outliers regardless of links to former topics. In this case, the outliers are not detected by missing links but by tell-tale features of innovative topics. Or presented the other way around: Outlier detection is in theory a clustering algorithm, in which deviant documents/instances are not clustered together with the other instances. Chapter 6 and Section 9.4 explain how additional features enable the selection of deviating documents.

4.4 Summary

This chapter presented the most relevant concepts in Machine Learning for this thesis: The distinction between Training and Test Set is important for the evaluation for Chapter 9, wherein the parameters of the approach are defined, i.e. the approach is trained. It had to be considered that the datasets for calibration and evaluation need to be distinct but also comparable. For that reason, the later introduced datasets for the evaluation (see Sections 9.2.3 and 10.2) were not used in other contexts in this thesis, while e.g. the datasets for the analysis of the bibliometric features (Chapter 6), the citation rate (Chap-

ter 8) and the calibration (Section 9.3) were in large parts identical. Recall and Precision were introduced as measures for coverage and accuracy of retrieval algorithms and their adaptation for the evaluation of clustering approaches in this thesis was shown. Other metrics like the Entropy and the Information Gain were explained for a better understanding of the rule derivation (Section 9.4).

The measuring of distance and similarity between vectors or clusters was described, as the second step of the system links topic clusters in different time periods according to their relatedness. Section 4.3 showed the necessary adaptations of the respective similarity measures for this thesis. The following chapter explains the concept of LDA and its adaptation for this thesis, so that the theoretic background covers all aspects of the system.

5 Latent Dirichlet Allocation (LDA)

5.1 Basic Approach

The underlying approach for the topic detection is based on an implementation of LDA (Blei, Ng and Jordan 2003, Heinrich 2008, Steyvers and Griffiths 2007). LDA deduces the topics in text documents by calculating the latent topic distribution of their words. Thus, each word belongs to a topic with a certain probability. The assumption that thus each document in turn is a composition of single probability distributions is based on the concept of mixture models in statistics. Before going into details for mixture models and their computation, the intuitive notion underlying LDA will be sketched.

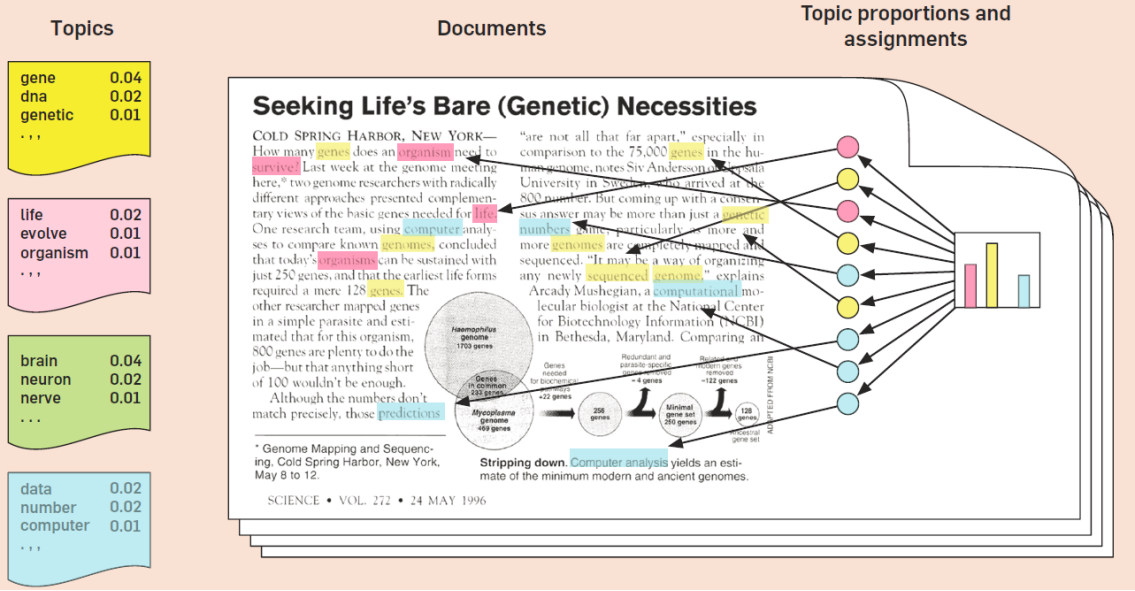
Each document is assumed to consist of a mixture of various topics, which are present to varying degrees. For instance, a document concerned with a countries flora might also cover related aspects like e.g. interaction with fauna, geography etc. Thus, the document is a mixture of topics, for which the exact definitions of the topics remain vague. Note that even for experts a clear delimitation of a topic's concept is difficult and only feasible in comparison with other topics.

LDA is called a topic mixture model because it uses distributions of words for distributions of topics. In general, any combination of underlying distributions can be called a mixture (Marin, Mengersen and Robert 2005). A mixture model consists of M observations, in this case the text documents, which in turn contain V categorical observations, the words, which are in the following called items. The points of observation contain information about parameters of the distribution indirectly (Marin, Mengersen and Robert 2005). In particular, information about the underlying probability distributions is only expressed latently. Mixture models assume that for each item there is a probability distribution for the K topics. According to this distribution, a probability distribution of mixture components for the K topics can be derived, which basically is the eponymous mixture or combination of distributions. Thus, the derived parameters majorly depend on the provided information.

The innovative aspect of LDA and similar approaches was the introduction of the so called topic mixture model as a contrast to clustering approaches. Even in fuzzy clustering, where a document is partially assigned to K topics, the document as a whole is associated with all these topics. For a better differentiation between both aspects, the fuzzy clustering can be imagined as a cluster or topic centred approach while the topic mixture models work on the document level. Thus, in fuzzy clustering a cluster might consist of documents where some are closer to its centroid and others farer away. This leads to overlaps in clusters where clusters might "share" documents to a certain degree. However, a mixture model regards the single items, i.e. terms, of each document and their respective probability distribution for the K clusters or topics.

In the case of mixture topic models, rather instead of a document being assigned to multiple topics the topics are associated with the terms and therefore the documents with a certain probability. Based on the observations, the probabilities of each item (the mixture components) can be calculated iteratively. They can be derived from the words used in the documents. Each word can represent and thus be mapped to each topic with a certain probability (which can nevertheless also be zero). Note that since the topics are assumed to be latent, the only observed variables are the words in the documents.

Figure 18: The idea behind LDA.



Source: Blei (2012, p. 78)

Figure 18 is an illustration by Blei (2012, p. 78), one of the inventors of LDA. On the left hand side are the K topics with their associated term probabilities.[62] On the very right hand side, the topic proportions $\theta(d,k)$ are depicted. They denote the likelihood for a topic k_i to appear in document d_j. According to these proportions, a topic k_i is drawn for each word. This resembles sampling with replacement, since probabilities for each topic stay the same over the whole process. Therefore, in theory, a document could contain only words from a single topic. Each topic k_i in turn has a word probability in the distribution $\varphi(k,w)$. Accordingly, a word is drawn from this distribution and placed in the text. This comprises the theory that enables the underlying assumptions about the text generation process. Because, if a topic is generated by these probability distributions, it should in turn be possible to approximate these distributions on the basis of the text documents. That is, if a topic appears in 60% of the documents, this topic's proportion should account for approximately 60% (if the number of observations M is large enough).

Figure 19 shows the general procedure of the approach. The initial setting assumes an equal probability for each term to appear in each topic and each topic to appear in a document (distributions φ and θ). However, in the beginning, the words are assigned randomly to the topics. This refers to each instance of a word, i.e. the term w appearing in document d on position n, and not a word in general. The result is the topic assignment $z(d,n)$ for an observed word $w(d,n)$. $w(d,n)$ is the actual value of a term in document d on position n. $z(d,n)$ is the topic that is associated with the word in document d on position n. The combination results in $p(w(d,n) \mid z(d,n), \beta(k))$, i.e. the probability of observing word $w(d,n)$ given the topic of the document and the distribution of the words on the topic.

62 In practice, each term appears in each topic but might have a probability close or equal to 0. Thus, each term probability vector has the same length for all topics. This is important when comparing these vectors later on e.g. with the Cosine similarity.

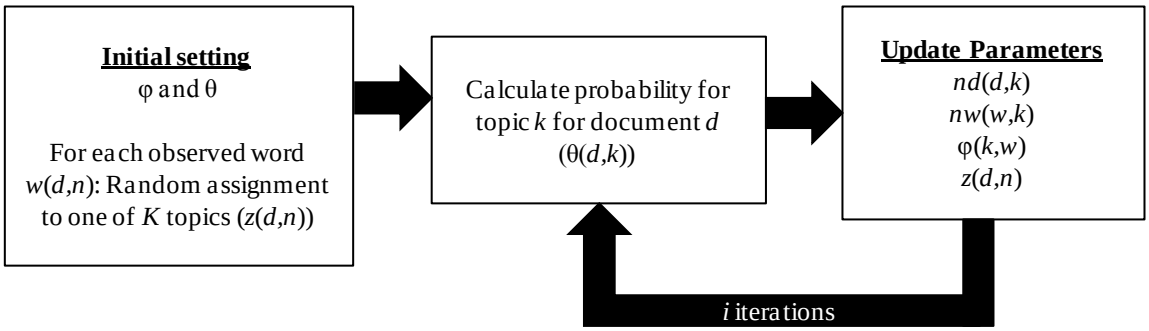
After that, the topic probability for a document can be deduced based on the number of words from a topic that appear in it and the word distribution for the topics (which is in the first iteration equal for all topics but will change later and is checked in the first iteration as in every subsequent iteration). In the first iteration, the topic probabilities for all documents are calculated based on $p(w(d,n) \mid z(d,n), \beta(k))$ given θ. Thus, for each document the topic probabilities are recalculated based on the current word to topic distribution but regardless of its current assignment. The probability of a topic *k* for a document *d* is therefore calculated as

$$\theta(d,k) = \frac{nd(d,k)+\alpha}{\sum_k nd(d,k)+K*\alpha} \quad (20)$$

where *nd* is the number of occurrences of terms assigned to topic *k* in document *d*. Then, the documents are reassigned in the topics that they fit the best.

θ represents the per-document topic proportion and is a symmetric Dirichlet distribution Dir(α_1, ..., α_k). Thus, its initial as well as all following values are dependent on α.

Figure 19: Iterations of the LDA approach.



Source: Own illustration

Notes: *w*(*d*,*n*) = term in document *d* on position *n*, *z*(*d*,*n*) = topic (for the term) in document *d* on position *n*, θ(*d*,*k*) = probability of topic *k* for document *d*, *nd*(*d*,*k*) = number of occurrences of terms assigned to topic *k* in document *d*, *nw(w,k)* = number of times word *w* is assigned to topic *k*, φ(*k*,*w*) = distribution of word *w* for topic *k*.

The per-document topic proportion θ(*d*,*k*) represents the composition of these single proportions. After calculating the topic-proportions, all other parameters can be updated accordingly. This includes the per-topic word distribution φ which is calculated as

$$\varphi(k,w) = \frac{nw(w,k)+\beta}{\sum_w nw(w,k)+V*\beta} \quad (21)$$

where *nw(w,k)* is the number of times the word *w* is assigned to topic *k*.

φ represents the probability of a word to occur in a topic *k*. This in reality latent information is derived from the distribution of words in the topics. As stated above, it depends on an initial random assignment. However, the process of assigning words to topics and topics to documents is repeated so many iterations *i* that this does not affect the result. It has been shown that LDA converges after *i*=10,000 iterations (Griffiths and Steyvers 2004), i.e. changes become less and less and the initial distribution of documents has no influence on the final result.

With this new φ, the word topic assignment *z*(*d*,*n*) can be recalculated for the next iteration. The new assignment depends on the probability of a topic appearing in the document and then this topic causing word *w* to be drawn. Thus, the formula can be rewritten to:

$$p(k|d,n) = p(w|k) * p(k|d)$$

$$= \varphi(k,w) * \theta(d,k)$$

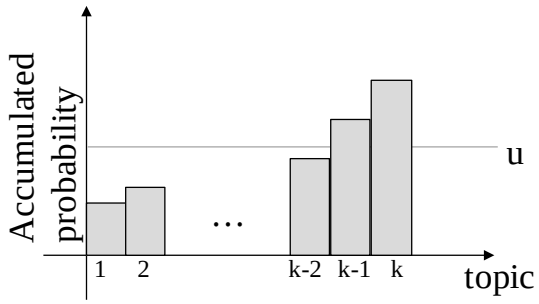
$$= \frac{nw(w,k)+\beta}{\sum_w nw(w,k)+V*\beta} * \frac{nd(d,k)+\alpha}{\sum_k nd(d,k)+K*\alpha} \quad (22)$$

The final assignment of a topic to *z*(*d*,*n*) is conducted via a scaled sample. Thus, the probabilities for the *K* topics are recalculated as

$$p(k) = p(k-1) \text{ for } k > 1 \quad (23)$$

and a random number *u* between 0 and *p*(*K*) is drawn. The topic *k* with the smallest value of *k* for which $u < p(k)$ is selected for *z*(*d*,*n*). Figure 20 illustrates this procedure: topic *k-1* is selected because it is the first topic – when ordered from the left according to the cumulated probabilities – for which $u < p(k-1)$. The higher the non-cumulated probability of a topic *k*, the higher the chance that *u* falls in the range [*p*(*k*-1); *p*(*k*)) since basically a bigger "impact area" is offered. However, the scaled sample enables to simulate drawing the topic rather than strictly assigning it.

Figure 20: The multinomial sampling with a cumulative method.



Source: Own illustration

In the end, the following estimations can be derived:

- The probability of each term *w* for a topic *k*: φ (*k*,*w*)
- The probability for a topic *k* to appear in a document *d*: θ (*d*,*k*)

Thus, for each topic, the set of important vocabulary can be derived by looking at the terms with the highest probability, i.e. those that are used most frequently in the respective documents.

Also, for each document the topic(s), to which it is most probably related, can be calculated. In order to achieve this, the probability for each document to belong to any of the *K* topics is calculated. Then, each document is assigned to the topic for which it has the highest probability. The topic with the highest probability for a document will be denoted as its "dominant topic" for easier references in the remainder of this thesis.

All in all, some parameters of the basic LDA algorithm enable an adjustment to the input parameters. In particular, these are the parameters α and β, the number of iterations *i* and the number of topics *K*. All these parameters stay fixed over the whole runtime of the algorithm and determine the initial values for θ (since the respective Dirichlet distribution is based on α), *z* (the multinomial distribution that is in turn based on θ) and φ (the distribution for the word probabilities that is based on β). The only observed variable that is unchanged during the whole process is $w(d,n)$, the true value of the word in document *d* on position *n*. Derived variables are the number of unique words *V* and the number of documents *M*. In the following, the input parameters are briefly explained in terms of their role for the LDA algorithm. Their final values for this thesis are set in Section 9.2.

α determines the shape of the Dirichlet distribution for the topics and thus the likelihood of a topic being selected. The exchangeable Dirichlet distribution, which is used in LDA, requires all elements of α being equal, leading to a set of topics that initially have the same likelihood. This avoids favouring any of the *K* topics since no prior knowledge about the topics is given.

The smaller α, the sparser is the topic distribution θ. A small value of α results in only few topics being associated with a document and thus the more α converges to zero, the fewer topics are assigned to one document at a time (with a minimum of a 1:1 mapping).

β is the equivalent for α in the multinomial distribution for the word-topic distribution φ. Again, it holds that the smaller the initial value of β, the sparser the distribution of the words on the topics.[63]

The integer *K* determines the number of topics that LDA generates. Thus, for a given *K*, in the first iteration of LDA, *K* topics are created and the *M* documents are randomly distributed among these topics. Also, it sets the size of the vector α, i.e. $\alpha = \alpha_1, ..., \alpha_K$, which is a uniform distribution anyway.

5.2 Adaptations for this Thesis

Preprocessing

Before applying LDA, a preprocessing step is conducted to eliminate stopwords and frequent words and to stem the remaining words. A list of stopwords was used to identify and remove common words with no content, as e.g. “and”, “with”. This list was supplemented with words that appeared frequently in the abstracts of the scientific articles as e.g. “Copyright”.[64]

Stemming was performed using Porter stemmer (Porter 1980).[65] Porter stemmer was designed for English texts and reduces all words to their word stem. No thesaurus is checked in the background. Instead, the endings of a word are stemmed according to a fixed set of rules. This might be problem-

---

63 To be precise, in the original definition, β is the initial probability for $p(w_j = 1|z_i = 1)$ for all combinations of $i \in K$ and $j \in V$ (cf. Blei, Ng and Jordan 2003).

64 Such annotations were wrongly but oftentimes included in the abstracts by the database providers. More such terms were found in a later dataset, which made further adjustment necessary (see Section 10.2.1).

65 http://tartarus.org/martin/PorterStemmer/, last accessed on 2013/03/13.

atic in cases in which the verb forms end with “y”, because Porter stemmer cannot detect whether such a word derives from a noun, adjective or verb.[66] This is certainly a drawback but the independence from any thesaurus or other hard coded part of the stemmer makes up for that. However, the main point is that words, for which the word stem is stable despite inflection, are reduced to the same word stem. For better illustration, a list of examples for Porter stemmer follows (non-existent words in italics):[67]

- liar → lie
- lie → lie
- lied → li
- lies → li
- *ly* → *ly*
- lying → ly
- *natur* → *natur*
- natural → natur
- naturalize → natur
- naturally → natur
- nature → natur
- natured → natur
- *natures* → *natur*

Terms (and references, see below) that appeared in too many documents were sorted out before applying the extended LDA approach. Therefore, after stemming and stopword removal, the term vector is used to determine the frequency of all words. Frequency is measured in document frequency (number of different documents in which a term appears) and term frequency (number of times a term appears in any document in total). This functionality is based on a notion of Blei in a lecture on LDA, in which he stated that common words should be eliminated before its application.[68] The criterion for the exclusion of common words can be set dynamically via the parameter t_w. The threshold t_w denotes the maximum share of documents in which a term can appear before being marked as ambiguous. The threshold t_w is relative to the number of documents in the set, M, and thus a value in the range (0,1]. All terms that appear in more than t_w*M documents are excluded from the further analysis.

For instance, a threshold of $t_w = 0.5$ corresponds to the fact that a word can appear at most in 50% of the documents. All terms marked as ambiguous are excluded from the LDA approach as they have no distinctive value for the topic model. On a similar notion, all terms that appear in only one document are excluded as the goal was to determine the representative words for a topic.

---

66 In other words: No hard coded lexicon is used.

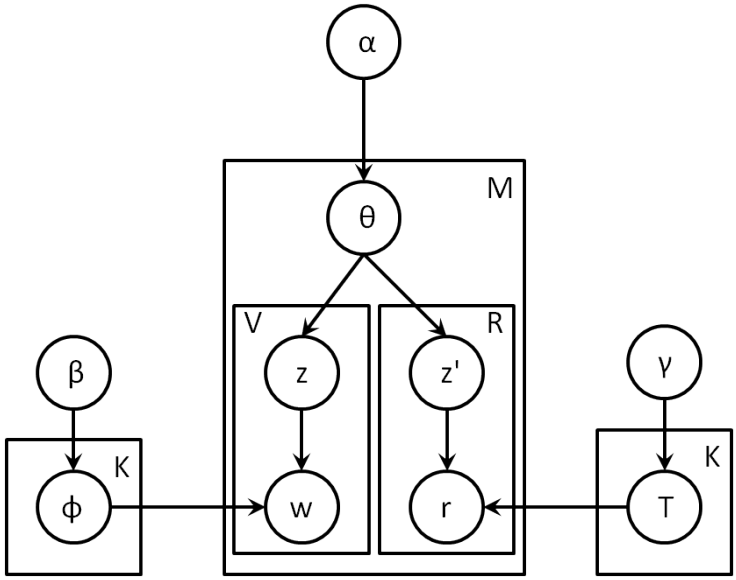
67 All examples taken from http://tartarus.org/martin/PorterStemmer/voc.txt and http://tartarus.org/martin/PorterStemmer/output.txt, last accessed on 2013/11/29.

68 Machine Learning Summer School (MLSS), Cambridge 2009, for reference see http://videolectures.net/mlss09uk_blei_tm/, last accessed on 2013/11/29.

References

LDA was enhanced by a feature that uses the references of a document for the topic models. Similar extensions have been made for web links as well as references in scientific documents (Erosheva, Fienberg and Lafferty 2004, Nallapati et al. 2008). Basically, a similar distribution as for words is applied for references. Thus, the additional parameters T for a topic reference distribution and its parameter γ are established. γ is the equivalent of β for the reference distribution among the topics. Thus, γ determines how specific a reference is for one topic. Therefore corresponding to φ, T shows the probability of each reference belonging to a topic *1...K*. Accordingly, T is a multinomial distribution with the parameter γ. Also *z'* is introduced as the equivalence of the term-topic assignment *z*. Variable *R* denotes the number of unique references in the dataset. Unlike words though, each reference can only appear once in a document. Figure 21 shows the overall LDA approach which is the result of this extension.

Figure 21: LDA extended for the usage of references (right hand side).



Source: Own illustration, adapted from Blei, Ng and Jordan (2003, p. 997)[69]

There are approaches that do not assume the same underlying set of *K* topics for the references (or links) and the terms from the documents' texts.[70] Rather, they have two sets of topics that are represented by the references and the terms independently. Here, both distributions, references and terms, use the same set of underlying topics. Thus, the probability of a specific document to belong to one topic k_i is calculated based on the probability of its references and the probability of its terms to belong

69 In the original illustration by Blei, Ng and Jordan (2003, p. 997), the multinomial distribution φ as described above is not explicitly shown. Instead, *w* is directly derived from β_k (via an implicit multinomial distribution). However, to better represent the actual implementation and calculations in the background, the illustration in this thesis also shows φ (and T respectively) as distribution for parameter β (γ respectively). Also, the labels of the parameters (in particular *M* and *V*) had to be adjusted in order to coincide with the remainder of this thesis.

70 See the "Alternative Models for References" by Erosheva, Fienberg and Lafferty (2004, p. 5233).

to topic k_i. Both probabilities are combined to get the overall probability of the document for topic k_i. The simplest way to do that would be to calculate the average of both metrics, resulting in a 1:1 weighting. However, a weighted combination of the probabilities takes into account that one of the two features might be a better indicator for a topic assignment than the other. Still, it would be possible to have an equal weighting for both metrics if it turns out that this is the best solution for the given scenario. Thus θ is now calculated as

$$\theta(d,k) = \frac{nd(d,k)+rd(d,k)+\alpha}{\sum_k(nd(d,k)+rd(d,k))+K*\alpha} \quad (24)$$

where *rd* is the number of occurrences of references assigned to topic *k* in document *d* (cf. Formula (20)) and

$$p(k|d,r) = p(r|k) * p(k|d)$$

$$= \mathrm{T}(k,r) * \theta(d,k)$$

$$= \frac{nr(r,k)+\gamma}{\sum_r nr(r,k)+R*\gamma} * \frac{rd(d,k)+\alpha}{\sum_k rd(d,k)+K*\alpha} \quad (25)$$

where *nr* is the number of occurrences of references assigned to topic *k* (cf. Formula (22)).

z' is calculated accordingly to z with a scaled sample. Since *z* and *z*' both determine the number of words and references respectively that are assigned to a document, they influence the values of *nd* and *rd* and thus the overall value of θ.

Usage of Multiple Textual Features

The standard LDA approach uses only one kind of text input. Some text documents, e.g. scientific publications or web pages, offer various kinds of text fields that can be used to extract textual features. In bibliometric databases, these include the title, keywords, abstract or the full text of a publication. In order to better take into account the different natures of the text fields, the LDA approach was extended so that not only the mere combination of two texts could be processed – which would correspond to attaching the text fields to one another – but also terms from different fields could have a different weighting. Therefore, the features F_t as well as their weighting w_t are defined.

The vector F_t determines which textual features of the bibliometric documents are used as input for the term based part of the LDA approach. Possible options with the datasets in this thesis are title, abstract, keywords and authors (which can be used as terms as well with certain probabilities to appear in each topic) and any combination of the aforementioned.

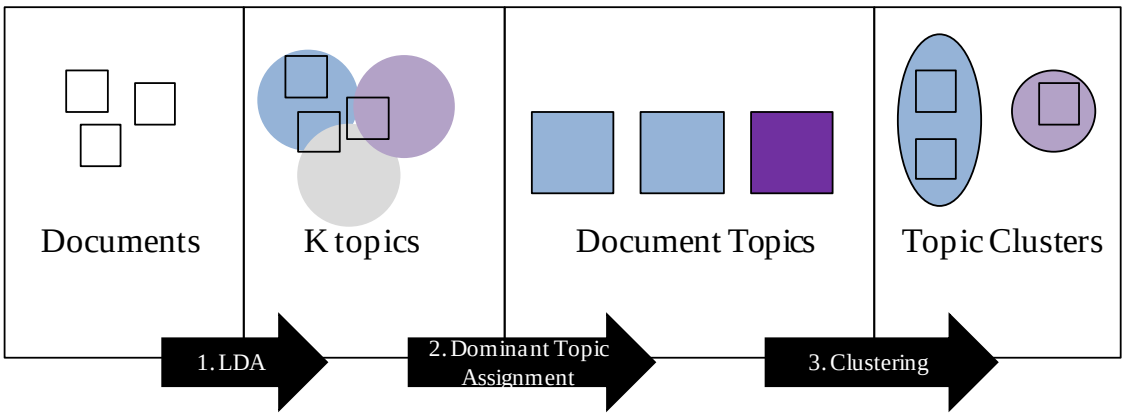
For a number of textual inputs F_t, the weighting w_t is an F_t-dimensional vector that associates a weight with an input. The weight determines how much a term of a specific input is valued in comparison with a term from another input source. The implementation is so simple that it explains this feature best. For a weight w_{ti}, the input data f_{ti} is read w_{ti} times. So, in the following it will seem as if a term appearing once in the input now appears w_{ti} times. To avoid unnecessary data, all values in w_t are divided by their greatest common divisor first.

In Section 9.2, different combinations of input data as well as varying values for w_t are tested, to derive the best setting for the textual basis and the importance of the single inputs for the topic model. The goal of this estimation was to see whether for instance the terms used in the title are more important than those in the abstract.

Dominant topics

Some evaluations made it necessary to have a 1:*M* mapping of documents to topics. This was in particular the case for experiments, in which the results of LDA had to be compared to a Gold Standard consisting of separate clusters. In order to calculate evaluation metrics correctly when comparing with such Gold Standards, the mixture model had to be reduced to a hard clustering. Thus, after LDA was run, each document is assigned to its dominant topic as its topic cluster. The term and reference probabilities generated by LDA are used to determine the dominant topic for each document. In this way, a (probabilistic) *K*:*M* assignment of topics to documents becomes a distinct 1:*M* association. Each document now only refers to the topic that it contains in the majority of its text and the references. When each document is assigned to only one topic, distinct topic clusters are created. These clusters can then be represented by a set of documents as well as their individual distribution of terms and references. Furthermore, they can be illustrated for instance by the top *n* most common terms.

Figure 22: Clustering via dominant topic estimation after topics have been defined.



Source: Own illustration

Figure 22 gives an overview of the steps from a mere assemblage of documents to the topic clusters. The first step is the topic generation via LDA. Based on the *K* topics defined in this step, the topic probability for each document is calculated. The dominant topic is assigned to each document to gain a distinct allocation of documents to topics which can be used in the end to group the documents according to their main topic.

Other text clustering algorithms would not additionally calculate these distributions but instantly create the topic clusters based on the similarities between the documents. This similarity could be based on the terms and references and other features, but the algorithm would not be capable to abstract from an equal weighting of the terms for different topics. For instance, if only the term vector was used, the clustering algorithm would merely compare the term vectors of two clusters, maybe even include the document frequency of the terms and thus account for overall common terms but not for terms that are

important for certain topics and not for others. Furthermore, if no abstraction from documents to the concept of topics is made, two documents might end up in the same cluster simply because they both share a number of words with a third document but not necessarily among themselves. Opposed to that, lifting the clustering on a topic level is a means for defining the vocabulary of a topic first before trying to fit specific documents in a particular shape.

Parameter *n*, variable parameters *K* and α

In this thesis, an additional parameter *n* is used to keep *K* flexible for varying numbers of documents *M*. *K* was not fixed because of the changing number of documents per year, but an additional parameter *n* was introduced that defines how many documents are assigned to a topic on average if there is a strict *K*:1 assignment. Thus, *n* denotes the number of documents a topic contains on average if the documents were equally distributed among these topics (which of course they are not) and $n = \frac{M}{K}$. More importantly in the case of this thesis, the number of topics to be built can be determined by $K = \frac{M}{n}$. The notion behind this supplementary parameter is to adapt the number of topics to the number of documents. The assumption is that larger document sets also contain more (latent) topics and vice versa. Also, it is assumed that the (average) size of the topics (or their spread) is independent from the specific year of analysis. Thus, while *K* is adapted to the varying number of documents, the number of documents per cluster should be rather stable.

This facilitated the usage of multiple datasets, e.g. when running the same instance of LDA on multiple years, because then *K* was automatically adjusted to the number of documents in each particular year. Thus, the problem of determining *K* so that it fitted for all numbers of documents in the various datasets was avoided by introducing parameter *n* which denoted the average size of the topic clusters. This parameter seemed far more flexible and especially far more intuitive because of the difficulty of deciding for a number *K* to denote how many topics can be expected in a set of *M* documents. In contrast to that, it is easier to estimate the size *n* of the desired topic clusters. A human end user can better imagine which granularity such a size would represent. Also this parameter adapts much better to varying numbers of documents that are fed to the LDA approach.

Similarly to that, α sometimes is set with respect to the value of *K*. A usual approach is to use $\alpha = \frac{50}{K}$, as has been done for example by Steyvers and Griffiths (2007) and Grün and Hornik (2011). A variant of this is the general setting of $\alpha = \frac{c}{K}$ with *c* being a constant (see e.g. Wallach 2008). Thus, as an alternative to a fixed value for α, the approach was extended to also compute $\alpha = \frac{50}{K}$. This also works with the variable *K* in which case $\alpha = \frac{50*n}{M}$

5.3 Summary

Table 6 lists all parameters that were mentioned in this chapter. In summary, there are three types of parameters: Those already incorporated in the original LDA approach, parameters that were added to allow the usage of additional data in this thesis and parameters that influence the connection of the resulting clusters. For the additional parameters, γ and w_r are the only parameters that directly concern the usage of references in the approach. All in all, there are in this context not many options how the

references can be applied. Thus, the only parameters left to estimate are the Dirichlet parameter and the relative weighting with regard to the term usage in the basic LDA approach. However, this should not diminish the importance of the extension. Both features, the wording as well as the references, can be influenced by the authors directly and are available as soon as the publication is published. The other parameters in the LDA extension, n, t_w and F_t are used to facilitate the usage of LDA for the publication data. The parameters in the last group are used to calculate the similarity between the clusters in different time periods. In theory, the respective implementation could built upon any form of clustering approach. The derived values for the parameters are, however, of course dependent on the similarity calculation which in the case of this thesis relies on the LDA topic models.

Table 6: List of parameters used in the approach.

Basic parameters LDA	**Description**
K	Number of topics created by LDA
α	Dirichlet parameter for topic-document association
β	Parameter for word-topic association
i	Number of iterations performed by LDA
Additional parameters	**Description**
n	Average number of documents per topic
γ	Parameter for reference-topic association
t_w	Relative threshold for word occurrences
F_t	Input used for term probabilities in LDA
Parameters connections	**Description**
t_c	Threshold for similarity between two topic clusters to establish a connection between them
w_r	The weight of the references in the similarity calculation
Y	Maximum or minimum time span between connected topics in years

Source: Own illustration

All of the parameters can be adapted to fit the data most efficiently. Section 9.2 explains how the values for these parameters were determined. However, having stated that i=10,000 is a sufficient number of iterations for convergences, this parameter is fixed beforehand.

III Contributions

6 Emerging Topics – What They Look Like

The purpose of this thesis is the development of a system that allows a facilitated detection of emerging topics in the vast amount of publication data. Their identification allows a better assessment of the current status of science but also of the individual documents. In Chapter 8, reasons against the usage of citation-based features for the detection of emerging topics are given. However, the therein presented results also suggest that the emerging topics are disadvantaged with regard to citation counts, so that the impact measurement of a topic or its documents should consider the development stage of a topic as well. In turn, the assessment of research topics according to their development stage can be used for different purposes, most importantly for decisions regarding the (financial) support of research groups and regions. For such objectives, it might also be necessary to decide whether the respective topics have the necessary prerequisites to persist.

In this chapter, the influencing factors of new scientific topics during their early development stage are determined. Documents in five pre-defined fields are analyzed with regard to the characteristics of the involved authors, their references and journals. With the help of an assignment to emerging and established topics, the publication behaviour of documents in different development stages can be compared. Foremost, indicators can be derived that can help to identify publications in emerging topics in science at an early-stage after publication as the features presented here are all available at the time of the publication of a document.

The chapter is structured as follows: An introduction to the work presented in this chapter and its relation to the thesis are provided in Section 6.1. In Section 6.2 the theoretical arguments for the specific characteristics of publications dealing with emerging topics in science are developed. Section 6.3 presents the data and describes the variables and methods used for the analyses. The descriptive and multivariate results are provided in Section 6.4. Finally, in Section 6.5 the implications of the findings with regard to this thesis are discussed.

6.1 Introduction

As has been discussed before, various reasons determine whether emerging topics might or might not establish themselves as independent research fields in the course of time (see van Dalen and Klamer 2005, Campanario 2009, Kilwein 1999, Benos et al. 2007). Besides structural factors like scientific and technological uncertainties, path-dependencies and lock-in effects (cf. Barber 1961, Johnson 2013, Stent 1972, Stent and Hook 2002), new findings are sometimes overlooked or rejected simply because the already established knowledge seems more intuitive or persuasive (Atkins 2003, p. 205) – a reaction that is not necessarily a result of the quality or potential of the finding itself (Kilwein 1999, Benos et al. 2007, van Raan 2004, Costas, Leeuwen and Raan 2011).

Chapter 1 already described how the scientific community ignored Mendel's work because of its innovativeness or "deviation" from established facts, patterns or methods (Atkins 2003, p. 47, Costas, Leeuwen and Raan 2011, van Raan 2004). In Chapter 2, the consequential necessity of pointers for publications in emerging topics was motivated. Therefore, this chapter was meant as a starting point to show in what way emerging topics differ in their publication behaviour from other, established topics.

The findings could on the one hand indicate hurdles in the publication process for emerging topics. On the other hand, a first impression of possible indicators for emerging topics can be gained. The features presented here were also a basis for the rules derived for the overall identification system in Chapter 9.

The goal therefore was to identify and test features that might help to distinguish publications dealing with emerging topics from the remainder of publications. A basic assumption is that the publication process for emerging topics is shaped by internal and external factors in such a way that it differs from the course taken by publications in more established topics. The associated characteristics for a publication are first and foremost deviations from the publishing "norm". They can be forced upon the publication if review or writing processes make it necessary to publish with certain co-authors from specific countries, in certain kinds of journals or with reference to specific former work (cf. MacRoberts and MacRoberts 1996). Even though these factors are "thrust upon" the publications and their respective authors by the circumstances, they can in turn help to identify such publications in emerging topics. Due to this "forced publication behaviour" it is in turn possible to detect the publications in emerging topics by these tell-tale characteristics. In the case that these assumptions about deviating publication characteristics and patterns can be substantiated, this might serve as a lever for the identification of emerging topics in science already at a very early stage after publication.

6.2 Theory & Hypotheses

Similar to the citations (cf. Chapter 8), other factors – more or less under the control of the respective author – prior to or following the publication process can be dependent on the acceptance of a topic. For instance, the development stage of a topic can influence its acceptance in (renowned) journals (Campanario 2009), the opportunities for collaboration, as well as the availability of the knowledge base and thus the references. Such features allow the indirect measurement of acceptance and spread of a topic. However, they are available at the time of the publication in contrast to the citation counts. Thus, the focus of this thesis lies on the identification of bibliometric indicators that are accessible ex-ante and are derived to allow for an identification of emerging topics in real-time. To be more precise, the aim is to apply exclusively bibliometric indicators that are available at the point in time when a paper is published. With the help of these indicators, documents in emerging topics could be identified, so that the respective topics can be deduced (or the documents clustered to represent these topics). For that purpose, in this chapter three groups of bibliometric indicators are differentiated, namely a) the features of the publishing journal, b) characteristics of the references of a publication and c) characteristics of the authors of a scientific document.

In this way, possible impeding as well as fostering influence factors regarding the publication source, possible influences of its knowledge foundation as well as underlying collaboration are analyzed. Thereby it can be deduced whether documents in emerging topics deviate in their bibliometric characteristics from those in established ones. This allows the inference of possible impediments or disruptive factors in the publication process for emerging topics. The findings show whether bibliometric indicators can be regarded as suitable indicators for emerging topics in the context of this thesis.

6.2.1 Journal-Specific Features

Journal size

When a topic is still relatively unknown, its importance and innovativeness might not be recognized. The placement of the topic to other existing topics in a later development stage might be difficult. Also, the acceptance and placement of the associated publications in journals can be aggravated.

Chew (1991) showed that rejected publications are often resubmitted to (and finally published in) journals with a smaller size and smaller circulation paths. Furthermore, smaller journals are by tendency more specialized, while journals with a higher page count have a broader focus (cf. Michels and Schmoch 2014). Thus, a resubmission in smaller journals can be seen as a process wherein a specialized document searches for its niche where it can be published. This might be a more important process for new emerging topics for which this niche has not yet been defined and thus is unclear for both the authors as well as the reviewers (cf. Thompson Klein 1996, p. 200).

Turoff and Hiltz (1982) stated that both journal size as well as rejection rate have increased, which augments the demand for highly specialized journals. This is especially due to the fact that the larger journals often have to perform a balancing act among various research areas. This bears implications on the readership as well as on the set of authors which are able to publish in these journals. Most importantly, new emerging topics are unlikely to fit into the concept of a journal that aims at a broad overview of ongoing research in many topics. A publications dealing with a highly innovative topic might thus be better recognized and acknowledged in a smaller, thematically more specialized journal.

Building on the work by Chew (1991), the size of a journal is measured by the number of articles it publishes per year. The basic assumption hereby is that the more articles a journal publishes per year, the broader its focus and vice versa. Good examples for journals with a broad focus are Science or Nature. In accordance with the findings of Chew (1991) and Ray, Berkwits and Davidoff (2000), it is analyzed if documents dealing with new emerging topics, for which publication in general might be more difficult, are published more often in smaller (more specialized) journals. This leads to the first hypothesis.

> *H1: The chance of finding an emerging topic in a journal is on average decreasing with the size of the journal.*

Journal age

The growth in scientific output and the spread of information makes the introduction of new, more specialized journals necessary (Turoff and Hiltz 1982). New journals are introduced when a topic becomes so important and established that a continuation of that topic in a separate community is foreseeable. For example, the journal “International Journal of Disaster Risk Reduction” was first issued in 2012 in response to rising attention to this topic and the need for a consensus on proper definitions in the field (Alexander 2012). However, it is evident that a topic needs to reach a certain level of establishment and dissemination so that the introduction of a new journal for it seems reasonable, i.e. the number of scientists involved and thus possible authors have reached a critical mass. Turning the ar-

gument the other way around, it is unlikely that emerging topics, i.e. topics for which only few (if any) publications already exist, can be found in newly issued journals. Therefore, documents published in younger, more recently issued journals should in the majority deal with established topics.[71]

H2: The chance of finding an emerging topic in a journal is on average increasing with the age of the journal.

Journal fields

One of the main sources for innovation is the combination of existing means and knowledge in a novel way. Exaptation, the misuse or adaptation of methods from other fields, is an illustrative example for innovation via combination (Johnson 2013, p. 172, cf. Section 2.1). Thus, topics in a still early development stage might be characterized by a higher interdisciplinarity.[72]

H3: The chance of finding an emerging topic in a journal is on average increasing with the number of fields a journal is classified in.

Journal Impact Factor (JIF)

Similarly to the size of a journal deducing its level of specialization, the prestige or standing of a journal can be represented by the JIF, i.e. the number of citations it achieved divided by the number of publications published in the respective years (see Section 3.2.3). Its value shows how much attention articles published in the respective journal receive. This in turn reflects the readership and shows the standing of a journal. As van Dalen and Klamer (2005) pointed out, the reputation of a journal can be seen as a signal for the scientific community.

It is questionable whether documents that start a new topic can be placed in journals with a high reputation, and therefore a high JIF, because the reviewers might be more critical[73] and the acceptance of new ideas lower. One of the reasons for rejections of papers can be "avoidance of unconventional ideas" (Benos et al. 2007, p. 147). The Nobel prize winners Hans Krebs and Barbara McClintock have been rejected in the journal Nature for their innovative papers (Kilwein 1999, Benos et al. 2007). Benos et al. (2007) conclude that that "avoidance of avant garde and controversial topics by reviewers and editors could hamper the advance of science" (Benos et al. 2007, p. 147). Reviewers can easier get by with such biased reviews if the journal has many submissions (cf. H1), which is one effect of a high standing or popularity. Thus, the following hypothesis is tested:

H4: The chance of finding an emerging topic in a journal is on average decreasing with an increasing JIF.

---

71 It should be noted in this context that Dirk (1999) was not able to show a relationship between the age of a journal and the originality of published articles.

72 More details on the notion of interdisciplinarity as an indicator for emerging topics are provided in Chapter 7.

73 Ragone et al. (2013) proved a high discrepancy between reviewer scores and citation counts for conference papers. Even though the notion of using citations as a sole quality indicator has its flaws (which the authors mention as well), the divergence cannot be denied.

6.2.2 Reference-Based Features

Age of references

In accordance with the saying “dwarfs standing on the shoulders of giants”, new (still small) topics might have to rely on more established, older topics. This relationship between a topic and “external” or former work can be traced by the references in the papers.

With external work, publications from other topics are denoted, which are used or adapted for the ongoing research. As Rinia et al. (2001) have shown, the citation delay for work from other disciplines is higher than for work from the same discipline. In other words, the knowledge transfer takes longer if disciplinary boundaries have to be crossed. As innovative research might make this necessary more often than traditional research, the age of the references is supposed to be higher.

Furthermore, more established, fundamental work in the own discipline might be used more often than ongoing research as the new discoveries can seldom refer to other current research issues. Thus, it can be assumed that the chance of finding a new document increases with the average age of its references.

H5: The chance of finding an emerging topic is on average increasing with the age of the references used in a document set.

Fields of references

The argumentation for the fields of the references is the same as the one for the fields of a journal. The question is just how the interdisciplinarity of a document is measured. In particular, with the requirements of this thesis, all metrics that use citations for the documents are excluded (like e.g. applied by Porter and Chubin 1985 in a similar context). Morillo, Bordons and Gomez (2001) showed that the measurement of interdisciplinarity via journal categories and reference categories does not necessarily lead to similar results. Therefore, both metrics are tested as they are not only used to measure the abstract concept of interdisciplinarity but also indirectly that of the emergence of a topic.

The references can reflect the specialization of a topic or its interdisciplinarity. For this feature, the number of different fields of the references of a document was calculated. This should reflect on how many fields a document in a discipline or a topic relies on. Analogous to H1 and H3, the underlying assumption is that new emerging topics are less focused and thus have to use more fields in their references than other topics (cf. previous reference to a more in depth analysis of this assumption in Chapter 7).

H6: The chance of finding an emerging topic is on average increasing with the number of distinct scientific fields the references of a document are classified in.

6.2.3 Author-Specific Features

Number of Authors

Collaboration among scientists can be hindered by the early development stage of a topic. Scientists with new ideas might be reluctant to share their research and rather keep the idea and the expected reputation for themselves. This might in particular concern the communication across research groups as trust cannot be as easily established in a remote collaboration (Olson and Olson 2000, Rocco 1998, Handy 1995). According to Olson and Olson (2000), "Rocco's (1998) result [...] suggests that team members should travel to remote sites to engage in a team-building activity to engender lasting trust" (Olson and Olson 2000, p. 169). Trust might play a more important role in emerging topics, where the involved researchers might be more cautious with regard to premature disclosure of the findings (as they give up part of the control of that knowledge, cf. Zand 1972). Trust is especially important in risky situations (Olson and Olson 2000), in which the disclosure of new findings – which is not supported (or driven) by a project – can be categorized. Similar observations have been made for interdisciplinary collaboration (cf. Anholt, Stephen and Copes 2012).

In addition, knowledge exchange might be difficult in fields that are not yet properly defined. This does not only hinder the direct communication but also fostering supportive activities like exchange programs, projects etc. To be able to write a project proposal which is accepted, either a common understanding or many preliminary studies are necessary. In the case of sustainability science, it was found that collaboration was hindered by borders between regional clusters as the focused fields were different and thus "collecting, exchanging, and integrating diverse types of knowledge" were inhibited (Yarime, Takeda and Kajikawa 2008, p. 19).

Therefore, authors in emerging topics might be very conservative in respect to co-authorships. The impeded collaboration might reflect in smaller author groups. Consequently, the higher the number of authors, the smaller should be the chance for finding a new emerging topic.

H7: The chance of finding an emerging topic is on average increasing with a decreasing number of authors named on a document.

Number of countries involved

As was already discussed in H7, communication about new topics might be more difficult than in other topics due to missing definitions and foundations of research. In particular, the dissemination of a topic across borders might not be as good for emerging topics as for established ones. An emerging topic can be locally/geographically bound. Reasons for this might be that necessary preconditions for collaboration are more difficult to attain in international settings.

Parallels might be drawn to research in small, more secluded countries which are more dependent on their innovative results and thus less open for collaboration (cf. Roolaht 2012, Perry 2001). Continuing this line of thought, Roolaht's (2012) statements about small countries and enterprises might also be applied in this context, where actors "succeed through focused collaboration in international networks

and by using niche strategies" (Roolaht 2012, p. 35). However, this might be seen contradictory as the collaborations might threaten the stand alone criteria and thus the settling in an open niche.

In this study, international spread of a topic is measured by the number of different countries from which the authors of a paper originate. For high values of this feature, it can be assumed that the topic is already well known across borders. This might concern in particular already established topics, as they do not have the above mentioned problems of missing trust and common ground. Emerging topics on the other hand might suffer from these issues and thus their geographical spread might be rather small.

H8: The chance of finding an emerging topic on average decreases for an increasing number of distinct author countries named on a document.

Table 7 summarizes the individual features as well as their hypothesized effects.

Table 7: Summary of the hypothesized effects.

Hypothesis	Feature	Type	Relationship to finding an emerging topic
H1	Journal size	journal-specific	negative
H2	Journal age		positive
H3	Journal fields		positive
H4	JIF		negative
H5	Age of references	reference-based	positive
H6	Fields of references		positive
H7	Number of authors	author-specific	negative
H8	Number of countries		negative

Source: Own illustration

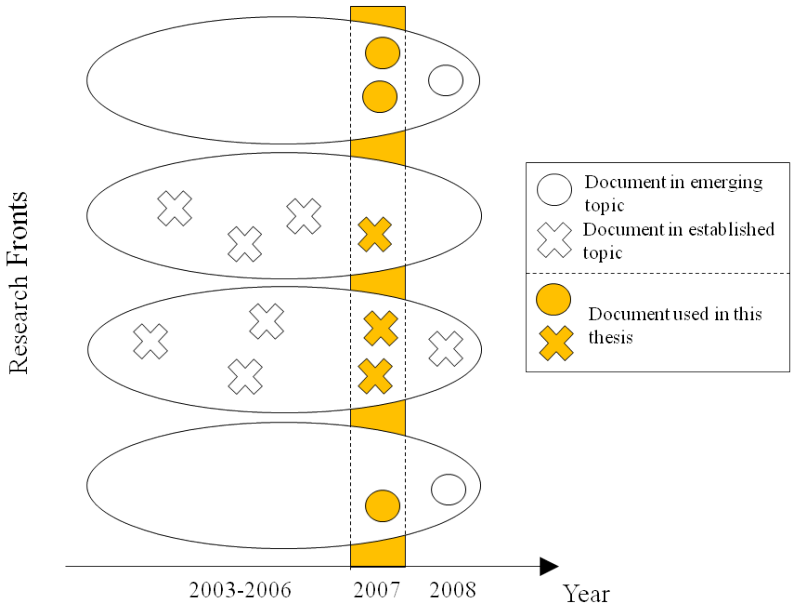
6.3 Data & Methodology

A necessary precondition for testing which of the features might help tell apart emerging from established topics is a dataset where a distinction between emerging and established topics has already been made. Therefore the classification of documents in emerging and established topics by the NISTEP is applied, which is based on an analysis of co-citations (cf. Saka, Igami and Kuwahara 2010). With the help of this classification, the explanatory power and significance of each of the discussed features can be tested with regard to identifying emerging topics as defined via the NISTEP analyses of co-citations. The features that provide significant explanatory power for the differentiation between emerging and established topics can subsequently be used as a stand-alone solution to separate emerging from established topics. This results in a set of indicators for the identification of emerging topics in science, which has the clear advantage of being available at a very early-stage after a document has been published.

6.3.1 The Data

The data employed for the analyses is based on a document collection created by NISTEP as a basis for a report on hot research topics (cf. Saka, Igami and Kuwahara 2010). In their work, Saka, Igami and Kuwahara (2010) collected the 1% most highly cited papers in 22 scientific fields for the years 2003 to 2008 from the WoS database by Thomson Reuters. Only journal publications from the Science Citation Index Expanded (SCIE) and the Social Science Citation Index (SSCI) were taken into account. The resulting documents (articles and reviews) were clustered thematically on the basis of a co-citation analysis to form what Saka, Igami and Kuwahara (2010) labelled “research fronts”. In a second step, Saka, Igami and Kuwahara (2010) identified emerging topics as those document clusters which only contained documents published in the years 2007 and 2008 that had no thematic overlap to any of the document clusters from the earlier years 2003 to 2006. The documents for the years 2007 and 2008, which had a thematic overlap to earlier document clusters were classified as "established" (Figure 23). For this study, this classification of established and emerging topics was employed, transferred to the document level and used to extract all documents (established and emerging) of the year 2007 from the WoS database.[74]

Figure 23: Identification of emerging topics by NISTEP and usage in this study.



Source: Own illustration, based on Saka, Igami and Kuwahara (2010)

This core dataset consists of 3,271 scientific papers including their assignment to one or more of 22 scientific disciplines and the information whether a document was classified as "established" or "emerging". Using it as a basis, additional data from the WoS database were collected, which are necessary prerequisites for the calculation of the features, e.g. the size of the journal in which an article

74 Due to the two year time-span between 2006 and 2008, the data for the year 2008 was excluded in this study to keep the potential errors as few as possible.

had been published and the number of authors per document. Since the number of documents defined as emerging was too small for statistical inference in a fine-grained field classification, the fields had to be aggregated to a more coarse-grained classification. 19 disciplines in the original dataset were therefore aggregated to form 5 disciplines for the further analyses.[75] It has to be noted that, even though this step was not avoidable in order to achieve a proper number of cases and statistically significant results, it also irrevocably led to a more imprecise analysis. However, this also avoided multicollinearity that could be caused by a too granular classification (cf. Wissmann, Toutenburg and Shalabh 2007).

Table 8 provides an overview of the distribution of documents in emerging and established topics by scientific fields.

Table 8: Overview of the distribution of documents in emerging and established topics by scientific fields.

Discipline	**Documents in established topics**	**Documents in emerging topics**	**Total**
Engineering	356	67	423
	84%	16%	
Biology, Environmental Science & Geoscience	679	56	735
	92%	8%	
Medicine	828	64	892
	93%	7%	
Chemistry	449	53	502
	89%	11%	
Physics, Mathematics & Computer Science	659	60	719
	92%	8%	
Total	2971	300	3271
	91%	9%	

Source: Saka, Igami and Kuwahara (2010), WoS, own calculations and illustrations

Clearly, the dataset has limitations as it contains only the 1% highest cited papers in 2007. Thus, there is already a pre-selection of papers based on citation values. This does not only restrict the dataset to a smaller size but also limits the generalizability of the results to a certain extent, as only publications that were deemed noteworthy by the scientific community are included. However, this in turn reduces the dataset to those emerging topics that have been considered important by the scientific community from the beginning. Therefore, whenever this dataset is used, the results apply to emerging topics that have already been labelled as successful from the start and are thus of particular interest for e.g. political decision makers.

75 The disciplines regarding the Social Sciences were dropped due to small numbers of observations even in an aggregated point of view.

6.3.2 Variables & Summary Statistics

In this subsection the calculation of the feature variables used for the further analyses is briefly discussed. The summary statistics for all variables are presented in Table 9 at the end of this subsection.

Following the theoretical discussion, the information if the document is classified as belonging to an "established" (coded as 0) or an "emerging" (coded as 1) topic within the NISTEP database is used as the dichotomous dependent variable (*newTopic*) for the regression models. The models are differentiated by the five science fields "Engineering", "Biology, Environmental Science & Geoscience", "Medicine", "Chemistry" and "Physics, Mathematics & Computer Science".

The features discussed in Section 6.2 will serve as explanatory variables. For all documents in the dataset, the following variables were calculated, which can – according to the theoretical discussion – be described as either journal-specific, reference-based or author-specific:

- Journal size
- Journal age
- JIF
- Journal fields
- Age of the references
- Number of fields of references
- Number of authors
- Number of countries of the authors

Basically, there are four features that can be regarded as journal specific. The *size of a journal* is measured by the number of articles published in a given journal in the respective year, in the case 2007. The *age of the journal* is defined as the number of years since its first appearance in the WoS database. In theory, the first appearance in the database and the actual appearance on the market can differ. In particular, this might occur due to a belated inclusion of a journal in the Science Citation Index (SCI). Since Thomson Reuters, however, demands the fulfilment of specific criteria of quality and quantity for the inclusion of a journal in the SCI, this can be seen as a certain selection mechanism for journals to reach a common standard.[76] The third journal specific measure is the *JIF*, which is a citation based indicator for the evaluation of the quality of a journal. It represents the prestige and reputation a journal has in the scientific community and is measured as the frequency with which articles in journals on average are cited in a given period of time (cf. Section 3.2.3). In this case, the JIF is defined as the number of citations in the year 2007 divided by the number of cited publications in the period 2005-2006. The final journal specific measure is the *journal fields* variable. Its calculation is based on the classification of scientific disciplines in WoS. It is measured as the distinct number of scientific disciplines, in which a journal is classified, and thus is used to assess the interdisciplinarity of a journal or conversely its level of specialization.

---

76 http://wokinfo.com/media/essay/journal_selection_essay-en.pdf, last accessed on 2013/06/25.

As for the reference based measures, information about cited references in a publication is used. This is first of all the average *age of the references* of a document. For that, the age of each reference is calculated as the difference between the year of the citation and the year of the cited publication. The values for all references of a paper are then averaged. Thus, the value indicates if the article in tendency relies rather on an older or newer knowledge base. Analogously, the number of *fields of the references* counts all the distinct science fields the references in the document are classified in based on the WoS classification of scientific disciplines (cf. Section 3.3.1). The frequency of the emergence of the respective fields is not taken into account. Therefore, a value of 5, for instance, might indicate 5 references to publications classified in 5 different, non-overlapping fields or a reference to one publication assigned to 5 different fields. In any case, this value indicates the field specific diversity on which a given document is building its knowledge base. This builds on the assumption that the interdisciplinarity can be derived from the disciplines of the references independently of their actual distribution among the respective documents.

The final set of features refers to the author specifities of a document. The first indicator calculated within this context is the *number of authors* that are named on a given document. This measure is based on the different standardized names of authors[77] in WoS and indicates the extent of collaboration that resulted in a given publication. In a similar fashion, the distinct *countries of origin of the authors* named on a publication are calculated in order to indicate if the publication is the outcome of a national or an international collaboration.

Table 9: Summary statistics.

Variable	Obs.	Mean	Std. Dev.	Min	Max
newTopic	3271	0.09	0.29	0.00	1.00
Engineering	3271	0.13	0.34	0.00	1.00
Biology, Environmental Science & Geoscience	3271	0.22	0.42	0.00	1.00
Medicine	3271	0.27	0.45	0.00	1.00
Chemistry	3271	0.15	0.36	0.00	1.00
Physics, Mathematics & Computer Science	3271	0.22	0.41	0.00	1.00
Journal size	3271	1478.57	1456.40	5.00	7266.00
Journal age	3271	35.42	28.15	0.00	105.00
JIF	3236	7.50	6.05	0.20	63.97
Journal fields	3271	1.72	1.20	1.00	7.00
Age of ref.	3271	5.18	2.05	0.00	15.17
Fields of ref.	3271	10.02	5.43	1.00	46.00
Nr. of authors	3271	8.57	31.94	1.00	1311.00
Nr. of author countries	3271	1.76	1.65	1.00	19.00

Source: WoS, own calculations and illustrations

77 Since this measure is based on standardized names, it might fail in the case of two co-authors of a single publication who share the first name initials and the last name. However, this can be considered as a rather rare event, which is why distortions on this indicator should be limited.

6.3.3 Estimation Method

For the multivariate analyses of the features and their power in explaining the variance between new and old research topics logistic regression models are employed as the outcome variable is dichotomous. In the logit model, the log odds of the outcome are modelled as a linear combination of the predictor variables (Long 1997). After estimating a general model including the publication features as independent variables and controlling for field specific effects, the models are re-run for each of the five scientific fields since it can be assumed that the explanatory power of the publication features varies across disciplines.

To interpret the model coefficients, marginal effects at the means of the independent variables were calculated. They reflect the probability for a publication to fall into the "newTopic" category as identified in the NISTEP dataset. More specifically, the marginal effect represents the effect of a one-unit change in the independent variable on the probability to belong to the *newTopic* category of the dependent variable (coded 1), holding all other variables constant (Long and Freese 2006). The interpretation of the coefficients is different for continuous and discrete independent variables. In the case of continuous independent variables, an infinitesimal change of the independent variable changes the probability to belong to the *newTopic* category of the dependent variable, i.e. that the dependent variable takes the value of 1, by X%. For dummy variables, a change of this variable from zero to one changes the probability that the dependent variable takes the given outcome value by X%.

6.4 Empirical Results

In this section, the results of the analyses will be presented. In a first step, a descriptive overview on the selected bibliometric feature variables is provided and it is explained how well they are able to discover differences between emerging and established topics. The second step will be to test the assumptions via multivariate regression models which provide us with a more detailed overview on possible combinations of features for the early-stage identification of emerging topics.

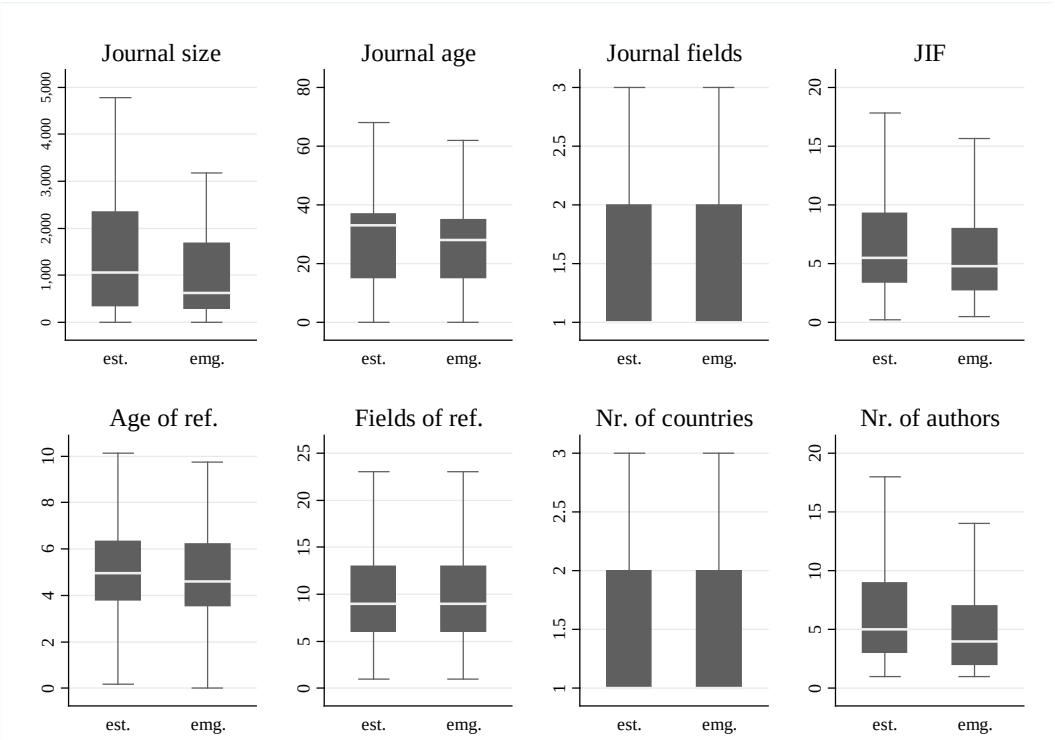
6.4.1 Descriptive Statistics – Box Plots

Before discussing the multivariate analyses, first of all a descriptive overview of the specific features and their relationship to established and emerging topics is presented with the help of box plots (Figure 24). The box plots show the distribution of the feature values in comparison. For the purpose of these graphs, outliers were excluded in order to allow for an easier visual comparison of the focal features of the distributions.

It is evident from the box plots that no overly large differences between the two types of topics regarding the single features can be found. The differences are largest for the journal size, where the median is smaller for emerging topics than for established topics, which is in line with H1. A similar observation can be made for the JIF (H4). Contrary to H2, however, the median for the journal age indicator is lower in the case of emerging than for the established topics. With regard to the reference-based indicators (H5 and H6) as well as for the number of authors (H7), only very slight differences between established and emerging topics can be observed. As for the journal fields (H3) as well as the number

of countries (H8), the variance in these variables is comparably low, i.e. 90% of the publications in the sample have a value of 3 or lower and a median of 1. Thus no differences between established and emerging topics for these two features can be found, at least not in this rather coarse-grained view.

Figure 24: Differences between established and emerging topics.

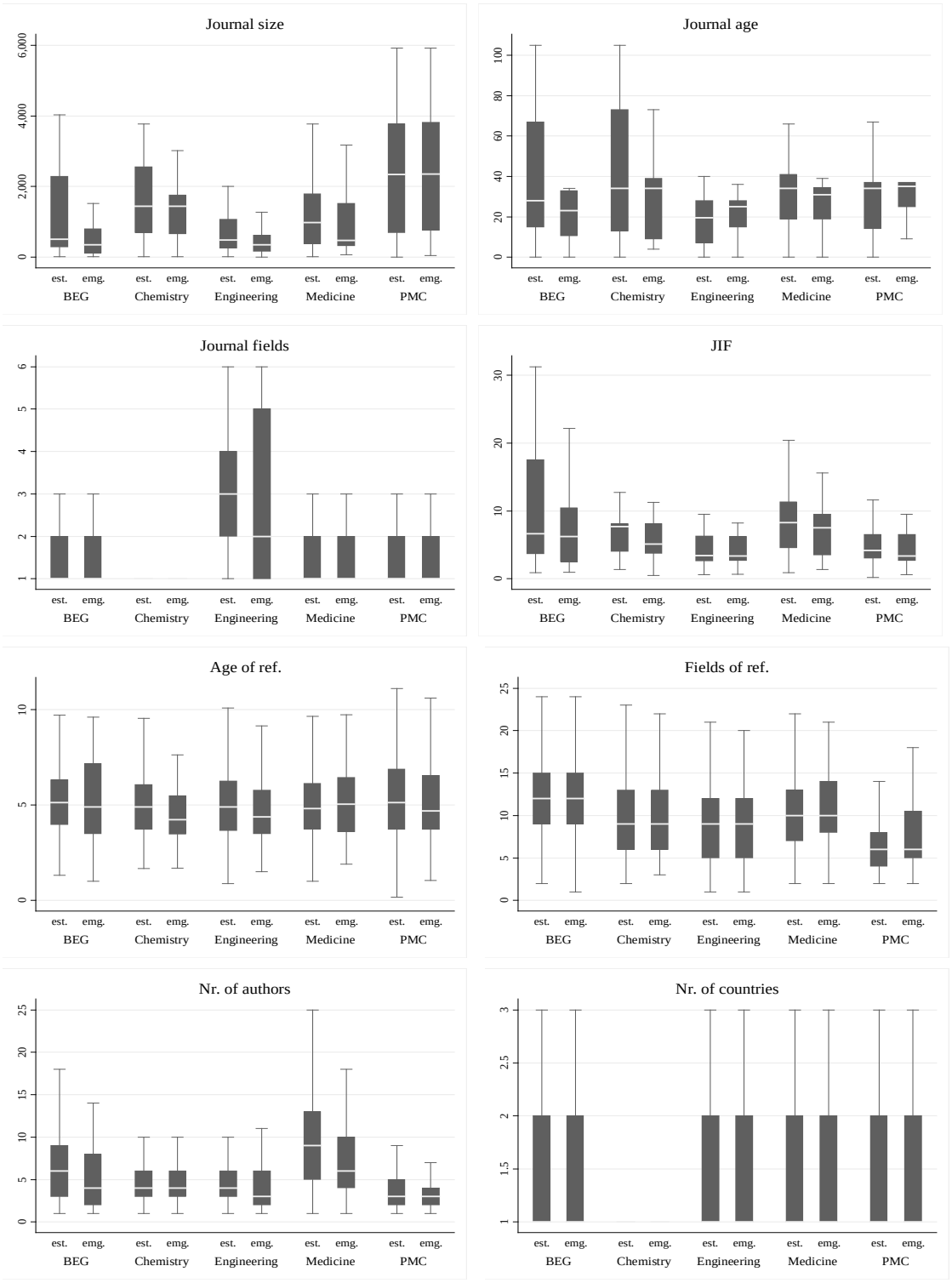


Source: WoS, own calculations and illustrations
Notes: est.=established, emg.=emerging.

When splitting up the box plots by scientific disciplines (Figure 25), the differences between established and emerging topics alongside the individual features become more evident. With regard to the journal size, which has been introduced as a measure of specialization, quite distinct distribution patterns can be found between emerging and established topics for all disciplines, except for Physics, Mathematics & Computer Science. As hypothesized, scientific articles dealing with emerging topics are generally published in smaller, more specialized journals (H1). In a similar vein, the feature on the number of fields of the journal can be interpreted, as it indicates, on the one hand, the level of specialization of the publishing journal, but can on the other hand also be seen as an estimate for the interdisciplinarity of the publication itself. However, the low variation on the variable does not leave much room for a descriptive interpretation, also when differentiating the results by scientific disciplines. Only for Engineering – which in general seems to have a higher interdisciplinarity than the other fields under analysis – differences between emerging and established topics become visible. Contrary to the hypothesis (H3), however, emerging topics in Engineering seem to be published less often in journals with a high interdisciplinarity.[78]

78 As for Chemistry, the distribution of the values is highly skewed. In 79% of the cases, journals are only assigned to one field. Thus, a box plot cannot be drawn in this specific case.

Figure 25: Differences between established and emerging topics across scientific disciplines.



Source: WoS, own calculations and illustrations

Notes: est.=established, emg.=emerging, BEG= Biology, Environmental Science & Geoscience, PMC= Physics, Mathematics & Computer Science.

With regard to the age of the publishing journal, the largest differences can be found in the fields of Medicine and Engineering. As for the JIF, a high variation in the values across the scientific disciplines can be found. In Engineering as well as Physics, Mathematics & Computer Science, for instance, the JIF is generally lower than in Medicine or Chemistry. Yet, in Chemistry, Medicine as well as Physics, Mathematics & Computer Science the highest differences in the JIF values are found between established and emerging topics. Across these fields, emerging topics seem to be more commonly published in journals with a lower JIF, which supports the hypothesis (H4).

For the reference-based measures (H5 & H6), i.e. the age and the number of fields of the references, a rather differentiated picture can be found, which does not allow for a general conclusion at this point. The range of the age of the references is similar across the disciplines, yet the differences between established and emerging fields are subject to high field-specific variations. As for the number of fields the citations are referring to, some variation across disciplines becomes evident. However, the differences between established and emerging topics seem to be minor across all fields.

Regarding the authors involved in a publication, on the other hand, differences between emerging and established topics are perceivable for all disciplines but Chemistry. In most disciplines, the number of authors per document is by trend smaller for emerging topics, which supports the hypothesis that collaboration partners are harder to find in a yet underdeveloped topic (H7). However, this observation does not reflect in the number of countries of the authors (H8), which is equally distributed for emerging and established topics, although this can once again be attributed to the rather small variance in this variable.

In sum, the evidence from the descriptive statistics points into the direction that an indicator system based on the selected bibliometric indicators may help with an early-stage identification of emerging topics in science. It is, however, already evident from the box plots that the fact that there are large variations across fields has to be acknowledged, implying that a simpler general model for the early-stage identification of emerging topics might not be suitable. Therefore, besides providing a general model, the multivariate analyses will be differentiated across fields, in order to additionally provide field-specific recommendations for the identification of emerging topics in science.

6.4.2 Multivariate Results

As a further step towards providing indicators for the early-stage identification of emerging topics in science, a set of logistic regression models was run with the discussed publication features as explanatory variables and the information if a document is dealing with an emerging or an established topic as the dependent variable. Analogous to the descriptive analyses, first of all a general model across all disciplines, yet controlling for the field specifities, was estimated. In order to test for multicollinearity between the explanatory variables, variance inflation factors (VIFs) were calculated based on an Ordinary Least Squares (OLS) model with *newTopic* as the dependent variable. As a rule of thumb, VIF values above 5 indicate a high multicollinearity between the variables. Besides the field dummies, which showed VIF values between 2.12 and 2.79, the journal age variable had the highest VIF (1.70). The mean VIF for the model was 1.67. Hence, no multicollinearity concerns can be found (O'Brien 2007).

Table 10: Logistic Regression – Marginal effects.

dV: newTopic	**dy/dx**		**S.E.**
Biology, Environmental Science & Geoscience	-0.059	***	0.018
Medicine	-0.064	***	0.018
Chemistry	-0.035	*	0.018
Physics, Mathematics & Computer Science	-0.057	***	0.017
Journal size	0.000		0.000
Journal age	0.000		0.000
JIF	-0.003	**	0.001
Journal fields	-0.005		0.005
Age of ref.	-0.004		0.003
Fields of ref.	0.001		0.001
Nr. of authors	0.000		0.000
Nr. of author countries	-0.010	*	0.005
Number of obs.	3236		
Wald chi²	50.1		
Prob > chi²	0.000		
Pseudo R²	0.026		

Significance Level: ***p<0.01, **p<0.05, *p<0.1

Source: WoS, own calculations and illustrations

Notes: "Engineering" is the reference group for the field dummies. For dummy variables, dy/dx is for discrete change of dummy variable from 0 to 1. The number of observations is slightly lower in the model than in the summary statistics as there are some values missing on the JIF variable.

Table 10 shows the marginal effects for this model. Overall, only the features for number of countries of the authors and the JIF show significant effects. The coefficients for both variables are negative. It can thus be concluded that the probability for documents in emerging topics to be published decreases with a rising impact factor of a journal, which supports the arguments of Benos et al. (2007). In addition, the significantly negative coefficient for the number of author countries provides evidence that the probability of finding an emerging topic on average decreases for an increasing number of distinct author countries named on a publication, which is in correspondence with hypothesis H8. For all the other feature variables, however, no significant effects can be found in this general model across disciplines. Yet, highly significant coefficients can be observed for the field dummy variables, once again indicating that the differences between established and emerging topics are varying highly across disciplines. The individual features might therefore also have different effects depending on the field.

In a next step, the models were therefore re-run separated by disciplines in order to find out which of the feature variables show significant effects in which of the scientific fields. Table 11 shows the marginal effects for the five disciplines.

Table 11: Logistic regressions for the single disciplines – marginal effects.

dV: newTopic	**Engineering**		**Biology, environmental science & geoscience**		**Medicine**		**Chemistry**		**Physics, mathematics & computer science**	
	dy/dx	S.E.	dy/dx	S.E.	dy/dx	S.E.	dy/dx	S.E.	dy/dx	S.E.
Journal size[α]	-0.167 ***	0.041	-0.013	0.014	-0.004	0.008	-0.009	0.038	0.019 ***	0.006
Journal age[α]	4.712 ***	1.669	-0.460	0.546	-0.095	0.369	0.101	1.012	0.417	0.921
JIF	-0.003	0.004	-0.002	0.002	-0.001	0.001	-0.004	0.005	-0.006	0.004
Journal fields	-0.002	0.017	-0.015	0.012	-0.007	0.011	0.009	0.011	0.002	0.007
Age of ref.	-0.009	0.008	-0.005	0.006	0.001	0.005	-0.015 *	0.009	-0.002	0.005
Fields of ref.	0.001	0.004	-0.002	0.002	-0.001	0.002	0.001	0.003	0.006 ***	0.002
Nr. of authors	0.002 *	0.001	-0.002	0.002	-0.008 ***	0.002	0.004	0.007	-0.006	0.006
Nr. of author countries	-0.021	0.025	-0.015	0.013	0.006	0.007	0.008	0.026	-0.015	0.021
Number of obs.	420		730		886		500		700	
Wald chi²	22.94		13.89		20.82		5.84		26.79	
Prob > chi²	0.003		0.085		0.008		0.665		0.001	
Pseudo R²	0.061		0.049		0.050		0.020		0.063	

Significance Level: ***p<0.01, **p<0.05, *p<0.1

Source: WoS, own calculations and illustrations

Notes: [α] Coefficients and standard errors multiplied by 1,000 to make effects visible.

As was expected, different effects of the independent variables can be found across the disciplines. Most distinctively, there are no significant coefficients in the field of Biology, Environmental Science & Geoscience, implying that within this field the early-stage identification of emerging topics with the help of these indicators is not possible. Similarly, in the field of Medicine, only one significant coefficient can be observed, namely the number of authors. The negative value of the coefficient shows that within Medicine the chance of finding an emerging topic increases when fewer authors are named on a publication. In Medicine, collaboration might therefore be impeded (at least on a small scale) for emerging topics. As for Chemistry also only one coefficient can be found to be significant. Here, the age of the references proves to be a valid indicator of the novelty of a topic. However, an increase in the average reference age by 1 decreases the chance of finding an emerging topic by 1.5%. This observation contradicts the hypothesis (H6). Documents in emerging topics in Chemistry seem to rely by trend more on a more recent knowledge base.

When it comes to Physics, Mathematics & Computer Science as well as Engineering, the indicators show a more precise picture. In Physics, Mathematics & Computer Science, two coefficients are found to be significant, namely the journal size and the number of fields of the references. The journal size is positively related with the development stage of a topic in Physics, Mathematics & Computer Science. Although the effect is relatively small as the journal size was measured via the number of articles of a journal in a year, this contradicts H1. Furthermore, the number of fields of the references shows a positive coefficient within Physics, Mathematics & Computer Science. Thus, at least for this field, the assumption can be confirmed that different fields are combined in the generation of a new topic.

Finally, in Engineering three variables are found to be significant, implying that an early-stage indicator system based on the publication features works best within this field. Here, the journal size is negatively related to the *newTopic* variable, i.e. an increase in article numbers decreases the chance of finding an emerging topic in this field (16.7% per 1,000 articles). The age of a journal also has a positive effect on the chance of finding an emerging topic (H2). For each additional year in the age of a journal, the probability to find an emerging topic increases by 0.5%. In Engineering, emerging topics are thus on average published in older journals. In addition, the number of authors named on a publication shows a significantly positive coefficient. In Engineering, documents dealing with emerging topics are thus on average published more often by larger research teams.

In sum, it can be stated that there are early-stage indicators for the identification of emerging topics in science. However, at the expense of a very timely availability of these indicators, one has to deal with certain inaccuracies that cannot be fully controlled. Using the indicators at hand thus provides the possibility of making a certain pre-selection of documents that might – with a given probability – deal with an emerging topic in science. After this pre-selection via the discussed publication features, still a manual search or a further analysis with the help of citation indicators is indispensable in order to truly find out if a publication in fact deals with an emerging topic or not. As soon as a document is identified as such, it can further be used to search for other publications that are thematically related. In addition, the analysis shows that the publication characteristics of emerging topics vary widely across disciplines. Therefore, only field specific analyses should be performed with the indicators that have been identified as possessing a given explanatory power in differentiating emerging from established

topics. To be more precise, specific characteristics of emerging topics involve the journal size and age, the reference age and the fields and the number of authors. However, the exact parameter values differed for the analyzed disciplines. This also makes it rather difficult to make clear statements about the hypotheses. In general, H4 and H8 can be supported, i.e. journal size and the number of author countries named on a publication are negatively related to documents in emerging topics. As for the journal size (H1) and journal age (H2) variables, evidence for the hypotheses was found only in Engineering. Similarly, for the indicators on the number of the fields of the references (H6) supporting evidence can only be found for Physics, Mathematics & Computer Science, while for the number of authors named on a publication (H7) support could be found only in Medicine. With regard to the number of the fields a journal is classified in (H3) as well as the age of the references (H5), no evidence in support of the hypotheses was found.

6.5 Summary

Various features of scientific publications were tested in order to identify a set of early-stage indicators for new topics. As mentioned in the beginning, the respective set of indicators not necessarily aims to be complete. That means in particular: 1) There are certainly more characteristic features for new topics which were not covered in this study and 2) the application of the features does not necessarily yield a result set covering all new topics from a document set. Furthermore, the features are highly dependent on relative factors and comparison to other publications in a dataset. Nonetheless, they offer a useful insight in the development process of topics and their initial obstacles in the publication process.

Specific characteristics of new topics involved the journal size and age, the reference age and the fields and the number of authors. However, the exact parameter values differed for the analyzed disciplines. The most pronounced discipline was Engineering for which a smaller journal size, a higher journal age and a higher number of authors were identified as indicators for new topics. Thus, publications in new topics were by trend published in established and specialized journals but with a higher collaboration effort than usual. In contrast, Medicine was found to have a smaller number of authors, which corroborated the assumption that collaboration might be hindered in new topics. Thus, there were two effects apparent: In technical fields, more researchers, research groups and/or equipment were needed in order to promote a new topic while in other fields, communication and thus collaboration might be impeded by the novelty of a topic.

These features of emerging topics might facilitate their promotion. By a heightened awareness of particularly these topics, their development can be channelled in beneficent directions at an early stage. Foremost, strategic planning is enabled, which fosters the optimal allocation of funding, equipment and work force. The set of indicators presented here allows only for a pre-selection step of candidates for emerging topics. However, given the vast (increasing) amount of annual scientific publications, such a pre-selection could be a crucial step in research monitoring.

Yet, there are still some limitations to this study. First, it was limited by the small sample size, which only allows restricted general deductions. Because of that, the disciplines had to be aggregated more coarsely, which made the results less precise. With a larger dataset, more profound and universal re-

sults could be expected. Second, the restriction to the top 1% highly cited papers might induce a bias: the pre-selection covered only documents that were already acknowledged in the scientific community at least within a time span of one year. This led to a distinction between “established” and “emerging and important” topics instead of only “established” and “emerging”. More interesting, however, would be a study based on a general dataset of emerging and established topics without such a limitation. Since many of the features are targeted towards the hindered publication process of novel ideas, the results gained with such a dataset could also be expected to be more universally applicable and distinct.

Finally, especially the variables for the number of fields of a journal and the number of countries had a very small variance, which handicapped their evaluation. More profound variables would be needed to represent a similar notion on a more granular level.

7 Emerging Topics – Interdisciplinarity as one Indicator[79]

As was discussed in the beginning (see Section 2.2), there have been various attempts in the past to identify innovative or high impact topics in science semi or fully automatically. Many studies did this in retrospective and with the help of citations (see e.g. Boyack and Klavans 2010, Shibata et al. 2009c and Price 1965) – a method that demands a time span of at least two years to give the scientific community enough time to discover, react and cite the topic in question (cf. Section 1.2). In this case, the identification of an innovative topic relies – on its basis – fully on the "wisdom of the crowd", i.e. the ability of the fellow researchers to discover and communicate the novel findings. Section 2.2 already listed reasons against citations as a measure for emerging topics. In Chapter 8, a detailed analysis of the citation counts of emerging topics will be given. In this thesis, the goal was the identification of indicators that are independent of the number of citations a topic receives. The citation analysis can however also be used to corroborate findings in this thesis, as done in Chapter 9 and in the current chapter.

In this chapter, the interdisciplinarity is tested as an indicator for highly innovative topics. Interdisciplinarity was defined – quite loosely – as the combination of different fields or even topics in a field. The only other necessary condition is that the combination of the topics is a novelty, i.e. the topics should have developed independently until then. Even though differences between multi-, inter- and transdisciplinarity exist (see e.g. Russell, Wickson and Carew 2008) the implications hold for all three kinds of combinations of knowledge across former boundaries. Thus the actual location for the knowledge combination is neglected, which is the basic differing factor for all three definitions, and merely its existence is noted.

7.1 The Relationship between Emerging Topics and Interdisciplinarity

The assumption that interdisciplinarity might be used as an indicator for innovation derives originally from Kuhn's definition of paradigm shifts (Section 1.1, Kuhn 1973, pp. 64f). According to Kuhn (1973, pp. 71, 77), revolution in science happens when a crisis appears. Crises in turn are evoked when present theories and methods are no longer sufficient to explain (new) observations or to fulfil the current needs (e.g. in the case of scarce resources). The solution to a crisis, a paradigm shift, can only be achieved if new theories or methods are introduced. The easiest way to do so is to be open-minded to standards in other scientific disciplines, e.g. when "scientists adopt new instruments and look in new places" (Kuhn 1973, p. 111). Thus, the transfer or adoption of knowledge across boundaries can help to turn the corner in a crisis (Thompson Klein 2004). For example, genetic algorithms use the basic biological principles of recreation and evolutionary survival of the fittest to facilitate complex mathematical calculations. In general, science is gradually but slowly becoming more interdisciplinary (Porter and Rafols 2009).

The combination of knowledge in turn can result in independent topics or fields (see e.g. Shafique 2013, Alvargonzález 2011) that can evolve in an independent way. Sometimes, this might lead again

79 This chapter has been published in large parts in Michels (2013).

to a diminishing multidisciplinarity, which might “hinder tapping the full potential of research” in severe cases (Shafique 2013, p. 77).

Furthermore, some findings already suggest that an interdisciplinary approach has more impact than a monodisciplinary one. As was explained before, the impact (measured in whatever form) is a reasonable indicator for innovativeness. For instance, it has been shown that multi- or interdisciplinary work enhances the citedness (Leimu and Koricheva 2005a, Levitt and Thelwall 2008) or the success rate (Sigelman 2009) and thus the impact of a paper. Albright (2010) argues that according to Adams (2006), creativity is a product of “the convergence of knowledge, creative thinking, and motivation” (Albright 2010, p. 105) and since these factors are promoted by multidisciplinary work, multidisciplinarity leads to creativity which in turn causes innovation (cf. exaptation in Section 2.1).

For single cases of innovative topics the interdisciplinary roots have been already shown. If innovative topics imply interdisciplinarity, does interdisciplinarity indicate innovativeness?

In order to answer this question, a Test Set was created in the field of Artificial Intelligence in Computer Science. The documents in this set were aggregated automatically in topic clusters separately for each year. The interdisciplinarity of each topic cluster was measured with different calculation methods. To compare these methods as well as test the overall hypothesis, those clusters with a high interdisciplinarity were selected and evaluated manually for their innovativeness.

7.2 Data

The document set was extracted from the in-house implementation of Elsevier’s Scopus database (as described in Section 2.4). All articles in Artificial Intelligence (Scopus code “1702”) that had a title and an abstract, at least 5 references and at least 2 citations were collected and 1,000 documents per year were selected randomly. The restriction to the documents meeting a specific threshold value was necessary to ensure a certain quality level in the sample set. However, the restriction might also influence the coverage of innovative work, as Chapter 8 shows. For the purpose of this study that not necessarily needed a full coverage of emerging topics, the bias introduced by such a low threshold level should be manageable. Because of the lower data coverage in the years 2010 and later, the data analysis was restricted to the years 2000 to 2009.

A first intention was to also include knowledge transfer to/from other fields in Computer Science, for instance from the field “Human-Computer Interaction” to Artificial Intelligence and vice versa. However, only few documents in Artificial Intelligence are exclusively categorized in fields of Computer Science (see Table 12). Most of them have a second assignment to fields in other classes. Thus, if an Artificial Intelligence document was cited only by Artificial Intelligence documents approximately 95% of these citations would be deemed interdisciplinary since the respective documents are also assigned to another category. Because of this, only non-Computer Science citations were classified as interdisciplinary citations.

Table 12: Percentage of documents in Scopus in the category Artificial Intelligence that are only assigned to Computer Science categories.

Publication year	Mono-disciplinary artificial intelligence documents
2000	1.4%
2001	2.1%
2002	2.6%
2003	2.7%
2004	2.7%
2005	4.2%
2006	4.7%
2007	6.0%
2008	8.0%
2009	6.1%
2010	4.7%

Source: Scopus, own calculations and illustrations

In the following, Scopus codes are aggregated to the first two digits to represent disciplines, e.g. all publications with codes starting with "16" are subsumed under the label "Chemistry" and denoted with "16XX".

7.3 Methodology

7.3.1 Clustering

The LDA approach as described in Section 5.2 was used to cluster the documents. The desired number of topics was k=120 per year.[80] $\alpha = 50/k$ was fixed, too; β was set to 0.01 as in other comparable studies (see e.g. Nallapati et al. 2008), γ was set accordingly. The textual input, F_t, for the word-based part of LDA was the title and the abstract of each document. The references were represented by the document IDs used in Scopus.

7.3.2 Interdisciplinarity

Interdisciplinarity was measured via citations and references for the documents. Furthermore, two aggregation levels were differentiated, namely the actual combination of disciplines or the mere combination of topics within the same discipline. Thus, four different kinds of interdisciplinarity for a cluster k were evaluated:

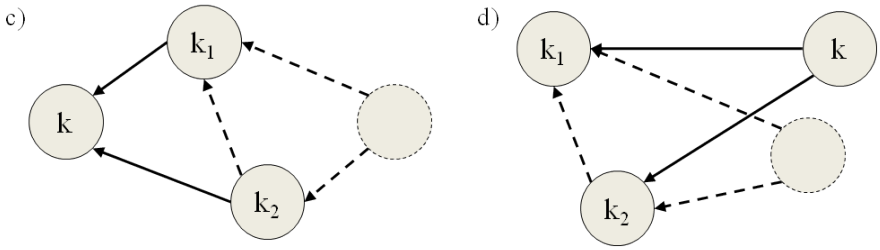
a) The percentage of cluster k's citations from documents in other scientific disciplines
b) The percentage of cluster k's references to documents in other scientific disciplines
c) The existence of other clusters citing cluster k for which the co-citation value is 0

80 As the training of the approach was conducted later (see Section 9.3), parameters had to be set according to literature and in such a way that the number of resulting topics was manageable. Therefore, the 1,000 documents per year were split between 120 topics.

d) The existence of other clusters for which the co-citation value would be 0 if not for the citation(s) by cluster *k*

Interdisciplinary work should be detectable either by its use of theories or approaches from other disciplines or by its applicability in other disciplines. Thus, metrics a) and b) were used to measure the interconnection with other disciplines. A topic cluster was deemed interdisciplinary if at least 50% of its references or citations were assigned to another (non-Computer Science) discipline. Even though probably not all topics come in the same form and can be detected by the same measure, all topics identified by the presented procedure can be deemed interdisciplinary.

Figure 26: The approach for detecting clusters that cite or are cited by different contexts.



Source: Own illustration

Notes: Arrows with continuous lines represent citations between clusters. The arrow points to the cited cluster. Dotted lines denote missing links via citations or missing linking clusters.

Not necessarily interdisciplinary in the usual sense, but still innovative are the clusters identified by the measures c) and d). These measures are included because the applicability of a topic in different contexts or a combination of different topics in a new context might be a sign of innovation. In this way, so far separate developments are combined. Figure 26 depicts these approaches and shows observed citations (arrows) and non-existent citations (dotted arrows and clusters).

Approach c) selects a cluster *k* if it is cited by two clusters k_1 and k_2 which do not cite each other and are not cited together by any other cluster. This metric is used to identify topics that can be applied in various contexts. Citations between clusters (for options c) and d)) are only taken into account if there are at least two citations. This condition avoids single appearances of connections. The threshold for crossing topical boundaries might not be as high as for connecting disciplines. But surely, this is the middle ground between building on former already intertwined work and genuine interdisciplinarity.

In approach d), a cluster *k* was selected if it cites two so far unconnected topics k_1 and k_2 (see Figure 26). Again, both clusters are not allowed to cite each other and be cited by other clusters conjointly. Since this led to many candidates for cluster *k*, a restriction on the time span of at least two years was introduced.[81] In this way, the time span between k_1 and k_2 and k_2 and *k* must be at least 2 years.

[81] See footnote 80.

7.4 Results

7.4.1 Fields

The interdisciplinarity across fields was measured with options a) and b). Table 13 lists the number of clusters in Artificial Intelligence for which more than 50% of the citations were emitted by documents in other fields (option a)). The cluster numbers are split up among those classes from which the respective citations came.

Chemistry, which had the highest count of clusters, targeted only clusters that contained misclassified documents in the Scopus database, i.e. documents for which the class Artificial Intelligence was assigned even though the document dealt purely with chemical aspects. The same holds for Social Sciences, Veterinary and Agriculture.

By contrast, citations from the disciplines Psychology, Neuroscience and Medicine indicate topics with computational as well as biological aspects, e.g. verb/sentence processing, visual perception and/or object recognition and learning in general.

Table 13: Number of clusters for which more than 50% of the citations stem from non Computer Science disciplines.

Scopus Code	Discipline	# of clusters cited by this discipline
16XX	Chemistry	22
32XX	Psychology	21
11XX	Agricultural and Biological Sciences	5
28XX	Neuroscience	3
22XX	Engineering	2
13XX	Biochemistry, Genetics and Molecular Biology	1
27XX	Medicine	1
33XX	Social Sciences	1
26XX	Mathematics	1
34XX	Veterinary	1

Source: Scopus, own illustration

Thus, in sum, there are two possible observations with option a): Either the topic indeed merges different disciplines or the documents in the topic were misclassified to Artificial Intelligence in Scopus. Both interpretations do not necessarily imply innovativeness.

Table 14: Number of clusters in which more than 50% of citations target a non Computer Science discipline.

Scopus Code	Discipline	# of clusters citing this discipline
16XX	Chemistry	24
32XX	Psychology	21
27XX	Medicine	14
28XX	Neuroscience	11
13XX	Biochemistry, Genetics and Molecular Biology	9
13XX	Biochemistry,	9
11XX	Agricultural and Biological Sciences	7
22XX	Engineering	3
19XX	Earth and Planetary Sciences	1
18XX	Decision Sciences	1
26XX	Mathematics	1
31XX	Physics and Astronomy	1
34XX	Veterinary	1

Source: Scopus, own illustration

The second variant of interdisciplinarity, option b), identified those documents that cited at least in 50% of the cases documents in other classes (Table 14). Again, a high citation rate of Chemistry etc. shows merely a misclassification of the respective documents. However, there are some topics that cite other disciplines extensively because they adopt or transfer knowledge.

A high percentage of these citations fall on similar aspects or rather aspects that are transferred from human behaviour to Artificial Intelligence, e.g. memory aspects. Even so, some topics can be found that introduce new techniques or applications, e.g. some clusters refer to Biochemistry or Medicine because Artificial Intelligence is applied to model biological processes.

Many clusters found with method a) are identified again, which indicates that applications in Artificial Intelligence that are based on psychological findings are also reused in the respective discipline. For instance, a topic that appeared in both lists deals with “Perception, distortion and degradation” and uses findings from Medicine and Neuroscience and is also cited in Neuroscience. Nonetheless approximately every second cluster is newly introduced in the list in approach b) and some of these topics seem a bit more focused in terms of content than those in the previous list. For instance, some topics deal with word ambiguity in text and not with speech processing in general.

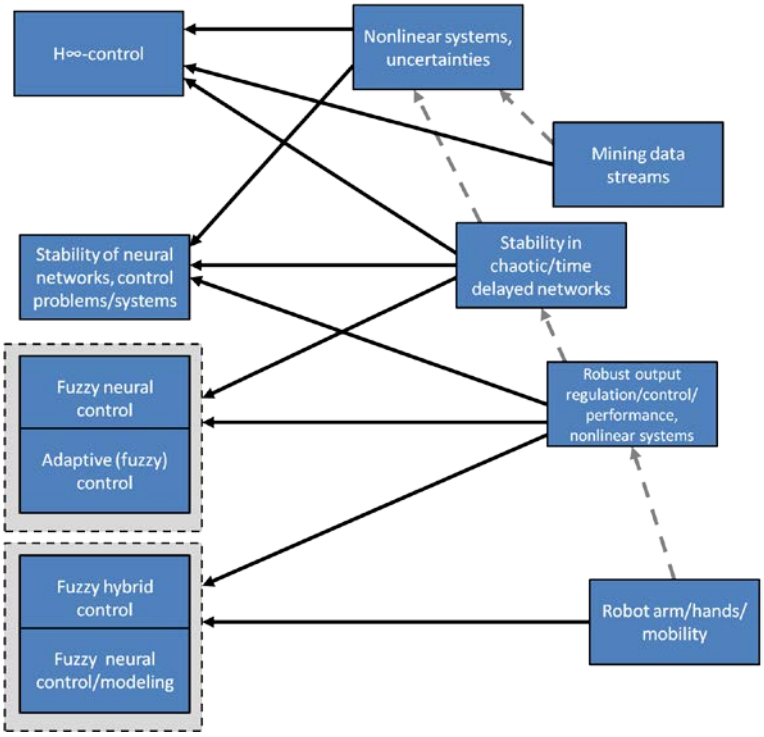
As the results of method b) seemed better than those of method a), it can be concluded that the interdisciplinarity of the reference list is better at indicating a high level of innovativeness than the citation list. This corroborates the findings in Chapter 6 for Physics, Mathematics & Computer Science. However, this indicator might be misleading as in cases shown above, which might be also a special characteristic of the field of Artificial Intelligence.

7.4.2 Contexts

As described above, the interweaving of different contexts was measured by options c) and d). The smaller number of identified clusters which even had a clearer focus makes it possible to present and discuss these results in more detail.

19 clusters were identified in total that were cited in different contexts (option c)), i.e. cited by clusters that were otherwise so far unconnected. Some clusters were detected by this method multiple times; for these clusters, different variants of triples *k*, k_1 and k_2 existed in which they held the position k but the clusters for k_1 or k_2 changed.

Figure 27: The network of 9 triples, in which the cited clusters connect yet unconnected clusters.



Source: Scopus, own illustration

Of the 26 different cluster triples (of the form *k* cited by k_1 and k_2) in total, 9 were interwoven in such a way that they can best be interpreted as a whole (see Figure 27). The clusters on the left hand side are those that are cited in different contexts (depicted on the right). Dashed arrows indicate that no citation between two clusters could be found even though they cite the same clusters. Thus, the notation mirrors the one in Figure 26. The grey boxes indicate clusters that were cited by the same clusters, e.g. in the case of “Fuzzy neural control” and “Adaptive (fuzzy) control”.

Most of the topics depicted on the left hand side of Figure 27 emerged before the observation period of the presented approach. Thus, they cannot be deemed as particularly innovative. For instance, the topics “Fuzzy control” and “H-infinity control” started well before 2000. Also, the network indicates that the topics in question had similar “siblings” and thus were not very innovative.

The remaining clusters identified by option c) are shown in Table 15. Clusters that appear in multiple triples are only shown once. The clusters represent rather ambiguous topics, that can be applied in multiple contexts, but which are not necessarily innovative.

Table 15: Clusters *k* cited by clusters (k_1 and k_2), which never cited the same topic before.

Clusters k		
Adaptive (fuzzy) control (2007)	Games, auctions, etc., optimization (2002)	Recommendation systems (2004)
Blind source/signal/component separation (2004)	Hierarchical structures with evolutionary algorithms (2004)	Source separation (2005)
Evolutionary/genetic algorithms (2002)	Image retrieval, recognition and classification (2004)	SVM (2001)
Face recognition/classification (2004)	Movement representation and perception (2001)	
Fuzzy classifiers, classification (2000)	Neural network performance optimization (2000)	

Source: Scopus, own calculations and illustrations

Notes: Only clusters not yet depicted in Figure 27 are shown. The clusters are presented in alphabetic order (in the order top to bottom, left to right).

The final analysis concerned those clusters that cited clusters which were so far unconnected (option d)). Table 16 shows the 6 triples of citing and cited clusters which correspond to *k* and k_1 and k_2 in the former description. The triples will be examined in detail in the following with a focus on the aspect of innovativeness of the citing cluster. For their discussion, the first column in the table introduces a numbering of the triples.

The citing cluster in the first triple shows best the idea behind the procedure in method d) – two so far unrelated topics (k_1 and k_2) are merged under a new aspect (in topic *k*). In this specific case, the findings in teleoperations, stability analysis and maintenance are aggregated in order to build robust nonlinear systems (Triple 1a). The same is done for fuzzy control and dynamic systems (1b). Their separate study forms a new research topic in 2009.

Table 16: Triples of cluster, where the citing cluster *k* connects two clusters (k_1 and k_2) that where never cited together before.

Triple No.	Cluster (k)	Cited cluster 1 (k_1)	Cited cluster 2 (k_2)
1a	"Robust" output regulation/ control/performance, nonlinear systems (2009)	Stability analysis	Teleoperations and autonomous vehicles
1b		Fuzzy/feed-forward control	Identification of dynamic systems
2	Sparse (Bayesian) modelling, dimensional reduction (2009)	Sampling/experiment selection for face/ object/image recognition	Independent component analysis
3	Classifier ensembles (2007)	Face recognition/ classification	Inconsistencies in structured text
4	Robot navigation, path planning (2008)	Target tracking with robots	Fuzzy behaviour based robots and multiple object tracking
5	Genetic algorithms (2008)	Dimensional knowledge reduction	Image transformation

Source: Scopus, own calculations and illustrations

Notes: For a better interpretation the publication years of cluster *k* are given in brackets.

Triple 2 shows a similar procedure in another topic. The initial cluster in the year 2009 combines different approaches to determine meaningful samples and models. All named aspects are not new themselves, but the approach of combining so far unrelated topics for the old task might be innovative. Analogously, neither face recognition nor text analysis are new topics (Triple 3), but the combination of their findings bears potential. Similar observations can be made for Triple 5.

In contrast to that, Triple 4 seems less innovative since both cited clusters are similar. It is noteworthy that both clusters – despite their similarity – are unconnected, i.e. have not been cited together so far or cited each other even though they seem to deal with the same topic. Still, their connection through the citations by cluster *k* is merely a result of their similarity and not an innovation process.

Table 17: Citation rates (average and maximum) for clusters that cite different contexts and for Artificial Intelligence in general.

Cluster	Average citation rate in cluster	Average citation rate in the same year	Maximum citation rate in cluster	Maximum citation rate in the same year
1	6.2	5.8	17	88
2	10.25	5.8	88	88
3	26.35	13.2	136	334
4	6.6	9.1	26	163
5	10.4	9.1	19	163

Source: Scopus, own calculations and illustrations

Despite the small yield of the approach, four out of the five selected clusters seem to be highly innovative. To evaluate their overall impact and potential, the citation rates for the respective documents were calculated and compared with the overall field value. Table 17 shows that the citing Clusters 2 and 3 have a citation higher than the average and Cluster 2 even contains the document in the dataset with the highest citation rate in the year 2009. Cluster 1, 2 and 5 have at least a citation rate higher than the average. This corroborates the findings suggesting that these clusters indeed have high potential in terms of innovativeness.

7.5 Summary

In previous work, the relationship between field dynamics, innovation and interdisciplinarity has been studied in one direction, i.e. it was shown that high innovative topics had a high interdisciplinarity. The results in this chapter suggest that the topics that were cited by different other topics (Method c)) were highly volatile, ambiguous or dynamic but not necessarily innovative. Being cited by other fields did not necessarily indicate a high innovativeness as well (Method a)). But the samples for clusters that cited other fields (Method b)) or differing topics (Method d)) extensively showed that indeed these clusters were in most cases high impact clusters. This level of innovativeness could be confirmed by a manual assessment of the respective clusters as well as their citation rates.[82]

For an illustrative interpretation of the results in the context of this thesis, the analysis was restricted to the field of Artificial Intelligence. It would be very interesting to know if a repetition of the approach in other fields yields similar results or not. At least in Artificial Intelligence, an automatic detection of highly innovative clusters can be achieved by analyzing their reference list. The reliance on the reference list in contrast to citations even allows the study of recent publications. Otherwise, a time lag between the topics and the citing publications would be necessary. The reference list is used in various contexts in the remainder of this thesis: First, the references are used as textual input to calculate similarities between documents and topics. Second, the number of fields in the references, which was also already used in Chapter 6, is applied in the rules for the outlier detection (Chapter 9).

---

82 The study via citation rates is limited due to the reasons given in Section 2.2 as elaborated in the next chapter. However, as stated before, while citation rates are not an indicator for innovative work, they can be used to assess the findings of an approach identifying emerging topics.

8 Emerging Topics – Why Citation Analysis is not an Adequate Metric[83]

This chapter supports the claim that citation analysis is not a valid measure for emerging topic detection. It might seem odd to dwell on a particularly inappropriate metric. However, citation analysis is one of the most widely-used metrics in bibliometric analysis[84] in general and for the detection of novel "high-impact work" (Small and Upham 2009, Mann, Mimno and McCallum 2006, Boyack and Klavans 2010, Shibata et al. 2008, Small, Boyack and Klavans 2014, see also Section 2.2). Because of this, rectifying the underlying assumption that citation analysis is not a reliable metric for innovativeness has become one main part of this thesis.

In the following, this will be shown with the help of the dataset acquired from NISTEP (cf. Section 6.3.1). It needs to be mentioned that, as this dataset was generated based on a citation analysis in the first place, the distribution of citations in this dataset is skewed. More pointedly, the expected citations in this dataset are higher than on average, as the contained documents all belonged to the 1% highest cited papers in the respective publication years and fields. However, the detection of emerging topics in the dataset was conducted based on a comparison of clusters in different time periods. Thus, restricting the view to the pre-selected set, the emerging topics are not necessarily (and – as will be seen later – neither practically) the documents with the highest citation rate. If anything, the pre-selection of the documents based on their citation rate might even be enforcing the observed effects. This will be discussed in further detail at the end of this chapter, when the results are at hand.

8.1 Methodology

The dataset was based on the implementation of the NISTEP dataset as explained in Section 6.3.1. This time, the original disciplines were used and the following six disciplines were selected for the purpose of this citation analysis:

- Computer Science
- Engineering
- Molecular Biology & Genetics
- Pharmacology & Toxicology
- Physics
- Plant & Animal Science

The documents of the NISTEP dataset published in 2007 were matched with the WoS database to calculate the citation rate over time with the most recent data. All citations were counted, i.e. alleged self-citations were not excluded. Also, no filter for document types in the citing documents was used. The derived variables are *citTotal*, which shows the number of overall citations between 2007 and

---

83 Major parts of this section were submitted as a research paper (Michels under review).

84 Also, in accordance with the original definition of bibliometrics, there is not much else left that could be used. Bibliometrics seems to be one of the many fields in science that is – per definition – resilient to changes and new methods.

2012, and *cit0, ...cit5*, which reproduce the annual citation rate in the time period between the publication year until 5 years after publication (i.e. 0, 1,..., 5 year(s) after publication).

With regard to the time window, the year of the first citation (*firstyear*) and the year with the maximum number of citations (*maxyear*) were compared. The year with the most citations is of course relative to the individual reception of a publication, i.e. the absolute values for this maximum might vary heavily. However, it considers the point in time when the publication received the maximum attention. This measured how long it took the scientific community to acknowledge the publication (since the restriction of the dataset demands that it is a highly cited paper anyway it seems that it is at least "worthy" of that attention). In other words: The *maxyear* reflects how long the publication had been publicly available until the masses recognized it.

All first citations happened in the year of publication or 1 year later.[85] Since the dataset consisted of papers published in 2007 and the citation analysis was calculated at the end of 2008, these were the only options for the papers that were among the most cited publications.[86]

8.2 Hypotheses

As discussed in the introducing chapters of this thesis (see in particular Sections 1.1 to 2.2), the acceptance and especially reception and adaptation of new topics might influence the time and the number of citations they receive. More precisely, this reluctant acknowledgement should lessen the total number of citations as well as the citations received in a certain time window.

Furthermore, the total number of citations might be influenced because fewer people are involved with a new topic. Since the research community is smaller, there are at first fewer people who would be even able to cite the work. Thus, this effect could lead to a smaller citation number in total for the documents published in new topics. In particular, it should be observable that the number of citations is by trend lower for documents in new topics in contrast to those in established ones.

H9: The total citation number is smaller for documents in new topics.

Using the same reasoning, it can be argued for a delay for the point in time when an article receives its first citation as well as its maximum citation number. With a lower acceptance rate and fewer researchers involved, the year of the first citation might be later for documents in new topics. Additionally, or because of that, the year in which the maximum number of citations is achieved might be "delayed" in comparison with other publications in established topics. As the citations can be seen as a measure of attention or application in the current time frame (cf. Sections 2.2 and 3.2.1) the year in

---

85 With the exception of one publication that was cited in 2006, i.e. one year before its publication. This respective publication in Physics was excluded from the dataset.

86 Please note that this is caused by the two-year time window (in this case the years 2007 and 2008) for the identification of the emerging topics in the original dataset; if the variance in the citations should be higher the time span for the identification of newly developed topics would have to be increased as well. However, with a citation window of 4 years, the "emerging" topics for the year 2005 would be found in a report in the year 2008.

which a publication reaches the peak in citations could also be influenced by the novelty of its topic. Using this reasoning, the following hypotheses were tested:

H10 - A: Publications in new topics are by trend cited later than those in old topics.

H10 - B: The year in which a publication reaches its maximum annual citation rate is later for documents in new topics than for those in old topics.

8.3 Results

Overview

Table 18: Overview of the dataset.

Discipline	Documents in old topics	Documents in new topics	Total
Computer Science	46	7	53
Engineering	206	39	245
Molecular Biology & Genetics	124	13	137
Pharmacology & Toxicology	19	7	26
Physics	558	50	608
Plant & Animal Science	167	9	176
Total	1,120	125	1,245

Source: WoS, own calculations and illustrations

For a better understanding and interpretation of the following analyses, a description of the used data is provided first. Table 18 gives an overview of the number of documents for each discipline and development stage of the topic. The distribution of new and old topics (and the respective selection criterion) might be biased across the fields, so that the field variable was used as a control variable in all following regression models.

Table 19: Overview of distribution of variables in the dataset.

Variable	Obs	Mean	Std. Dev.	Min	Max
newTopic	1,245	0.1	0.3	0	1
citTotal	1,245	117.4	161.0	4	2,830
cit0	1,245	6.2	9.1	0	108
cit1	1,245	24.9	27.0	2	409
cit2	1,245	26.8	34.7	0	590
cit3	1,245	25.7	40.3	0	731
cit4	1,245	23.8	40.7	0	765
cit5	1,245	9.9	18.1	0	332
firstyear	1,245	0.3	0.4	0	1
maxyear	1,245	2.1	1.1	0	4

Source: WoS, own calculations and illustrations

The single variables and their distributions are listed in Table 19. The variable *newTopic* indicates whether a document was published in a new or an old topic (cf. Chapter 6). An entry of "1" in this variable corresponds to a publication in a new topic, "0" to a publication in an established topic. The mean of *newTopic* corroborates the statistics in Table 18: The number of publications in old topics is much higher. The variable *citTotal* represents the total number of citations a publication received in the observation period. This value corresponds to the sum over the variables *cit0* to *cit5*, which account for the annual citations in that time period. Since the dataset contains only cited publications, the minimum value of *citTotal* is necessarily greater than 0. *cit0* shows the citations in the year of publication, *cit1* the citations in the first year after publication etc.

The variable *firstyear* represents the year in which the publication was first cited. It was calculated by subtracting the publication year from the citation year. In theory, this feature could have values between 0 and 5, but in the available dataset, there were only the values 0 and 1 (and one case of -1 as discussed before), as only publications in 2007 were contained that had been cited by the end of the year 2008 (see above). Since the dataset contained only the highly-cited papers, (at least) all publications that were not cited by then were excluded. Thus, the variable resulted in a binary codification whether a publication was cited for the first time in the publication year (0) or in the first year after publication (1). This had to be acknowledged in the statistical models. The *maxyear* represents the year in which the publication received its maximum annual number of citations. If more years with the same maximum value were found, the earliest of them was used. Again, this is a relative value representing the years 0 to 4 after the publication year.

Table 20: Correlation between the variables.

	citTotal	**cit0**	**cit1**	**cit2**	**cit3**	**cit4**	**cit5**
cit0	0.5698*	1					
cit1	0.9385*	0.6584*	1				
cit2	0.9773*	0.5370*	0.9424*	1			
cit3	0.9883*	0.4978*	0.8981*	0.9571*	1		
cit4	0.9750*	0.4609*	0.8542*	0.9274*	0.9749*	1	
cit5	0.9412*	0.4070*	0.7961*	0.8828*	0.9472*	0.9687*	1
maxyear	0.2621*	-0,0553	0.0962*	0.2233*	0.2989*	0.3496*	0.3369*

Significance Level: *p<0.01

Source: WoS, own calculations and illustrations

Table 21: Pairwise polyserial/polychoric correlation for variables *firstyear* and *newTopic*.

Pairwise polyserial correlation				
	firstyear		**newTopic**	
	Rho	S.E.	Rho	S.E.
citTotal	-0.5478	0.0491	-0.4000	0.0708
cit0	-0.9916	0.0000	-0.7093	0.0201
cit1	-0.7028	0.0376	-0.5616	0.0437
cit2	-0.4532	0.0639	-0.3828	0.0749
cit3	-0.4238	0.0677	-0.2648	0.0794
cit4	-0.3821	0.0709	-0.2444	0.0834
cit5	-0.3231	0.0709	-0.1799	0.0795
maxyear	0.1164	0.0391	0.1612	0.0478
newTopic	0.5856	0.0459		

Source: WoS, own calculations and illustrations

Notes: In case of the correlation between the variables newTopic and firstyear, the correlation coefficient is polychoric rather than polyserial since both variables are categorical. In all other cases a polyserial correlation is calculated.

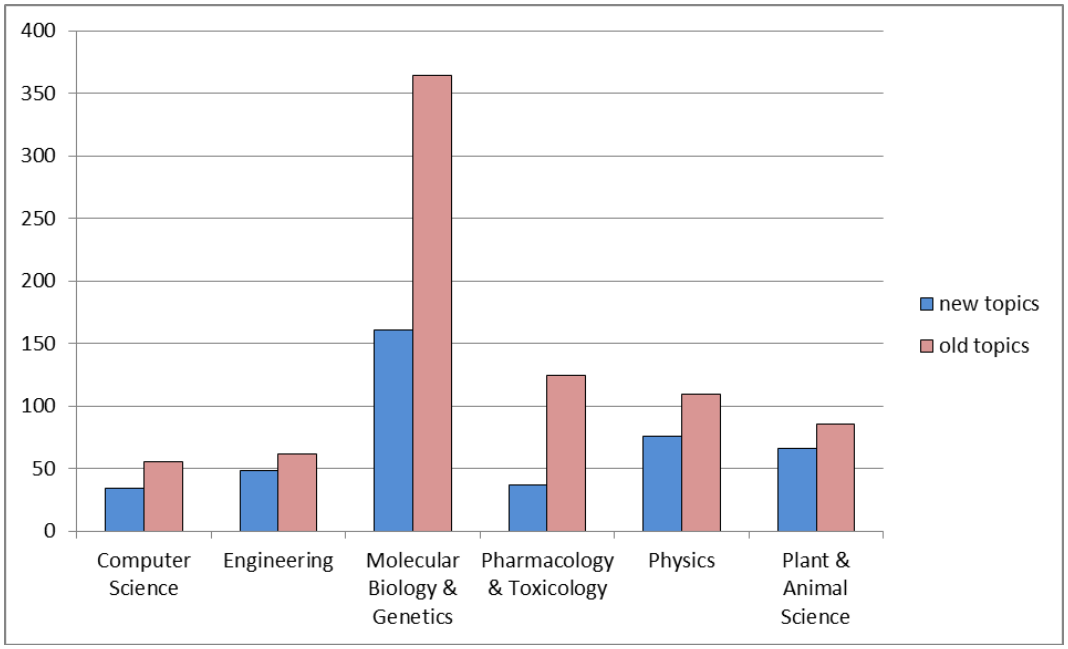
Next the pairwise correlation between these variables was analyzed (see Table 20 and Table 21). The variables *firstyear* and *newTopic* both can only take the values 0 or 1. Therefore, the correlation between these variables and the remaining ones was calculated as a polyserial correlation for which no significance levels are available. The values for Rho and the Standard Errors are given in Table 21.

Correlation with the total number of citations is the highest for the citations 3 years after publication. Note that a standard citation window of 3 years uses only the citations in *cit0* to *cit2* (like suggested for example by Glänzel and Schoepflin 1999). The variable *newTopic* shows a negative correlation with the total citation number, indicating that those documents with a value of 1 for *newTopic* have fewer citations than the others. The correlations of *citTotal* with the *firstyear* and the *maxyear* on the other hand seem rather tautological, as it is a natural effect that publications which are cited later are also cited less in a restricted time window. However, there is also a correlation between *firstyear* and *newTopic*, which supports hypothesis H10 - A, which stated that publications in new topics tend to be cited later than other topics. Similarly, there is a correlation between *maxyear* and *newTopic*, indicating that the latter could latently influence the citation process.

8.3.1 Citation Count

Figure 28 depicts the average number of citations per paper at the point of calculation. Numbers are shown for each discipline and each status of topic, i.e. old and new. The graphs for average citation numbers after 1 year and 2 years are nearly identical, only on other levels, and are thus not depicted here as well. The absolute values are relatively high – another side effect of using only the top 1% highly-cited papers.

Figure 28: Average citation rate in total.



Source: WoS, own calculations and illustrations

The ratio of new and old topics varies between 130% (Plant & Animal Science) and 341% (Pharmacology & Toxicology). The average number of citations for old topics in Molecular Biology & Genetics accounts for 227% of those for new topic publications. The latter two extreme cases indicate that progress in a topic might foster citations. Reasons for this might be that established topics are more commonly known and thus new publications are perceived by more scientists. Some journals and publication databases offer push-services which send an automatic message to the user if a publication that matches his defined keyword search is published. The necessary keywords should be easier to define for established topics.

The extreme discrepancies between new and old topics in Molecular Biology & Genetics and Pharmacology & Toxicology might be due to a wider spread of topics in those disciplines, where it is even more difficult to keep track of new findings if one is working in the field. Or it might be the case that new topics are more specialized and cannot be as easily adapted to other research.

Table 22 shows the results of a regression estimating the total citation number while controlling for the disciplines. As the dataset contains only documents that have been cited at least once in the observation period, a zero-truncated negative binomial regression model was used for the estimation of the citation counts. An ordinary negative binomial regression would try to predict zero counts despite the fact that there are no zero values in the observed data, which in turn would lead to biased estimates. Thus, the dependent variable in this model is the citation count as a non-zero integer number. The independent variables are the discipline variables as explained above as well as the binary coded status of the document, i.e. emerging topic (1) or not (0).

The regression confirms the hypothesis H9 and the findings in the pairwise correlation: The value of *newTopic* influences the total citation number negatively, i.e. for new topics the expected citations are significantly lower.

Table 22: Regression model for the total number of citations.

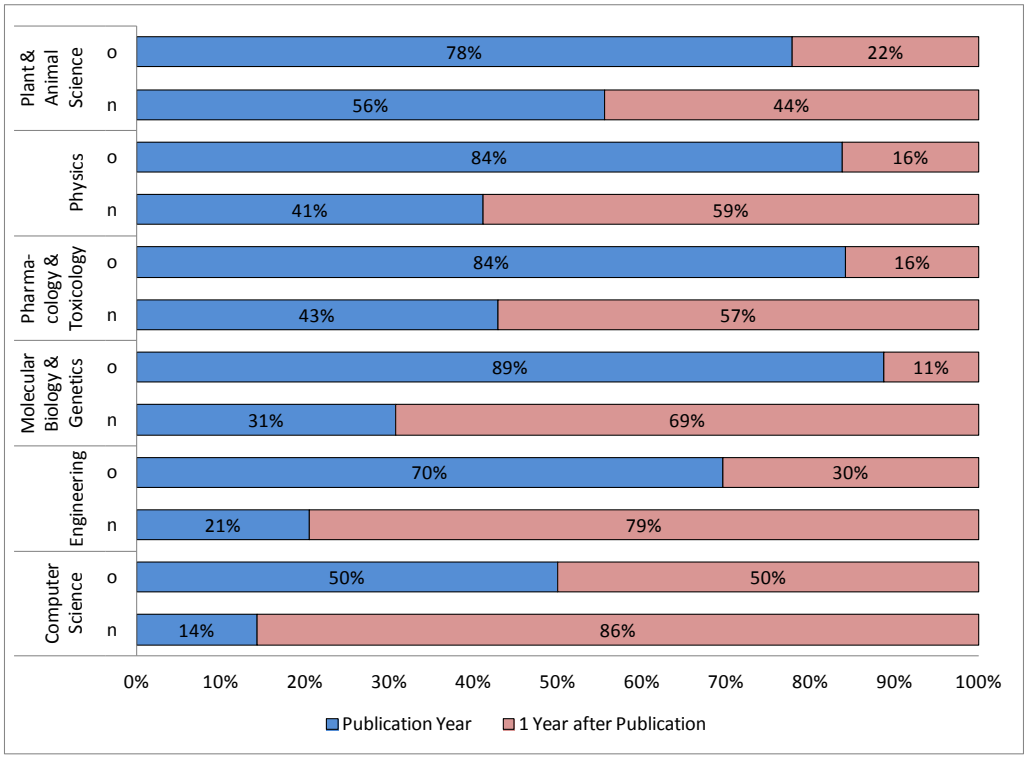
# Citations	**Zero-truncated negative binomial regression**		
	Coef.		Std. Err.
Engineering	0.146		0.104
Molecular Biology & Genetics	1.870	***	0.111
Pharmacology & Toxicology	0.666	***	0.163
Physics	0.691	***	0.098
Plant & Animal Science	0.456	***	0.107
newTopic	-0.399	***	0.065
Constant	0.098	***	0.026
Observations		1,245	
Pseudo R^2		0.047	
Likelihood ratio chi-square test		672.200	

*Significance Level: ***p<0.01, **p<0.05, *p<0.1*

Source: WoS, own calculations and illustrations

8.3.2 Time Window

Figure 29: Year of the first citation of a publication.



Source: WoS, own calculations and illustrations

Notes: For each discipline, new (n) and old (o) topics are analyzed separately.

Figure 29 shows a comparison of the years of first citations of publications in new (n) and in old (o) topics. In all disciplines, the documents in old topics are cited more often in the publication year than those in new topics. In Molecular Biology & Genetics, this accounts for 89% of the publications. In most other disciplines, this ranges from 78% to 84%. In Engineering, only 70% are cited in the publication year and in Computer Science, there is even a 50:50 chance of a paper being cited in the publication year or the first year. On the contrary, apparently the chance of a document being cited in the first year is higher for old topics. The disciplines in which this is most pronounced are again Engineering and Computer Science. This might be due to the fact that citations one year after the publication are more common here than in the other disciplines. However, the difference between the types of topics is observable throughout all disciplines.

Table 23: Regression model for the year of the first citation.

Firstyear	**Logistic regression**		
	Coef.		Std. Err.
Engineering	-0.804	**	0.319
Molecular Biology & Genetics	-1.887	***	0.375
Pharmacology & Toxicology	-1.683	***	0.571
Physics	-1.607	***	0.305
Plant & Animal Science	-1.313	***	0.340
newTopic	2.020	***	0.210
Constant	-0.015		0.286
Observations	1,245		
Pseudo R^2	0.109		
Likelihood ratio chi-square test	153.750		

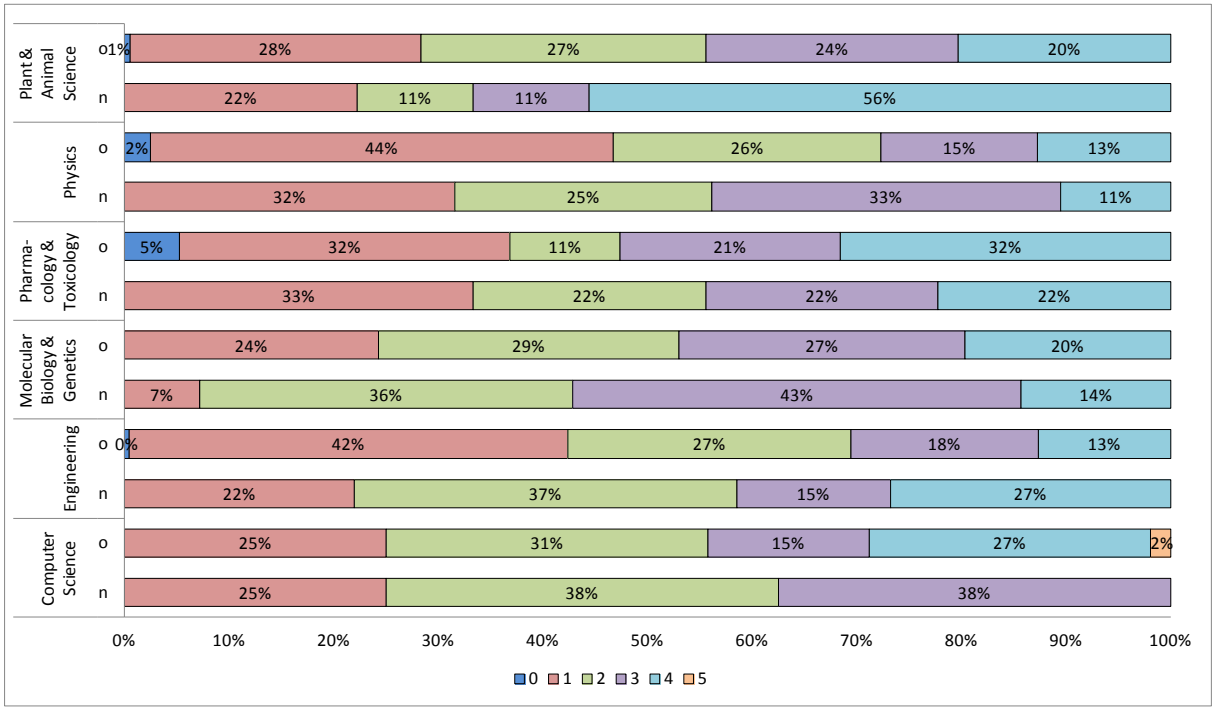
*Significance Level: ***p<0.01, **p<0.05, *p<0.1*

Source: WoS, own calculations and illustrations

To test the hypothesis that first citations are delayed for documents in new topics, a regression model was applied to look for dependencies between the variables *newTopic* and *firstyear*. Again, the independent variables were the control variables for the disciplines and the status of the document's topic. A logistic regression had to be used as the dataset only allowed for two options: Citation in the year of the publication or later (i.e. in 2008). Thus, the dependent variable was a binary codification of the year of the first citation and the model used a logistic model which calculated the probability of a citation in the publication year as a linear combination of the status of the document's topic and the scientific field.

Indeed, there is a positive significant influence of the value of *newTopic* on the first citation year (Table 23). Translating the binary codification of these variables, the first citation happens by trend in the year after publication for documents in new topics, while the other documents are cited in the publication year.

Figure 30: Number of years, after which the maximum number of citations was achieved.



Source: WoS, own calculations and illustrations

Notes: For each discipline, new (n) and old (o) topics are analyzed separately.

Next, the year, in which the maximum number of citations for each publication was achieved, was calculated (Figure 30). The year was again set relative to the publication year, so that values between 0 and 5 were observed. However, only few publications reached their peak in the first (0) or last (5) year of the observation period. Here the differences between the new and old topics are not as pronounced as they were before. Also, the observed trends are not similar across all disciplines. In Computer Science, there is almost no difference but only that more publications in new topics achieve their maximum citation count in the third and not the fourth year. In Molecular Biology & Genetics, Engineering and Physics, publications in old topics reach their peak more frequently in the first year. In Plant & Animal Science, the majority of documents in new topics clearly attain their maximum value in the fourth year while those in old topics are equally distributed over the first to fourth year.

Thus, no overall trend can be observed. However, it can be said that some publications in new topics would be disadvantaged if the citation window is too small. This is in particular the case for the new topics in the disciplines Plant & Animal Science, Engineering, Physics and Molecular Biology & Genetics if the citation window is less than four or three years respectively. Above all, if the analysis is reduced to the first year after publication, only 25% of the documents in new topics would be covered, while 38% of those in old topics would have already reached their peak. This might be important if the top cited documents are selected.[87]

87 In this case, the analysis was already performed on a dataset that was selected by exactly this method. The possible bias is discussed in the summary.

Table 24: Regression model for the maximum citation year.

Maxyear	**Negative binomial regression**		
	Coef.		Std. Err.
Engineering	-0.167	*	0.101
Molecular Biology & Genetics	0.026		0.105
Pharmacology & Toxicology	-0.018		0.157
Physics	-0.213	**	0.095
Plant & Animal Science	0.015		0.103
newTopic	0.145	**	0.063
Constant	0.830	***	0.090
Observations		1,245	
Pseudo R^2		0.008	
Likelihood ratio chi-square test		32.330	

*Significance Level: ***p<0.01, **p<0.05, *p<0.1*

Source: WoS, own calculations and illustrations

Despite the findings above, the regression model in Table 24 shows a significant influence of the variable *newTopic* on the maximum citation year. In this model, the maximum citation year was used as a count variable in a negative binomial regression. To account for the unobserved heterogeneity between observations in the data, the negative binomial regression uses an overdispersion parameter "alpha" (cf. Long and Freese 2006, p. 243).

The model shows that the peak in citations occurs later, if the document deals with a new topic. The Pseudo R^2 in all three models was highest in the logistic regression for the year of the first citation. As was discussed before, the number of citations depends on various factors (quality, visibility, etc.) and can thus not be explained as good as the actual year of the first citation. This variable is the one for which the available variables have the highest explanatory value.

8.4 Summary

The regression models showed that the development stage of a topic can influence the citations for the respective publications. In particular, the citation rate is significantly lower for publications in new topics. Depending on the scientific discipline, the average citation count for publications can be more than twice as high if it is written about an established instead of a new topic. Also there is a higher chance that it will be cited earlier than innovative work. Also the year in which a publication in a new topic reaches its citation peak can be delayed in comparison to other papers. The findings suggest that there is indeed a disadvantage for innovative work.

However, the reasons for this also lie in the nature of the research. The research community for new topics is smaller, thus the number of papers in which the work could be cited is lower. Also publication sources might be fewer and the full scope of the findings might not be acknowledgeable or usable for "outsiders" of the topic. Regardless of the reasons, the publications in new topics are disadvan-

taged in citation analysis. This might be a major issue when the results are used to evaluate performance (as innovative research is underrated) or in particular to identify “important” publications.

Limitations of this analysis are grounded in the dataset size for some categories (i.e. Pharmacology & Toxicology and Computer Science) as well as the dataset generation. The origin of the dataset was already mentioned in the beginning as based on highly cited papers in the disciplines. This might introduce a bias. Since the new topics are already those that are cited less in this dataset, the trends should be transferable to a general setting in which they should be even more pronounced. Given the time window, there were also some restrictions because only publications that were at least cited in 2008 were covered. In this way, only new topics that were identified by other researchers in the first two years were included, which also might reduce the differences found. In this case as well, a higher variation is to be expected if the analysis was to be performed on a more general dataset. However, the creation of such a dataset would be a very time consuming as well as subjective task and thus also prone to (other) biases. For this thesis, the important main outcome was the perceived bias that corroborated the assumption that apart from a time lag, other reasons limit the usage of citations for the detection of emerging topics.

9 Emerging Topics – How They can be Detected

This chapter shows the calibration and application of the approach – as proposed in Sections 4.3 and 5.2. After a summary of the overall approach is given, the Training and Test Sets are described (Section 9.2). Based on these datasets, the parameters for LDA and the similarity calculation as well as the rules are estimated. The resulting approach is then assessed in Section 9.5. In the end, a summary of the main observations is presented.

9.1 Proposed Approach

The system for emerging topic detection proposed in this thesis consists of three components (cf. Sections 1.2, 4.3 and 5.2):

- Topic clustering
- Topic connection
- Emerging topic selection

In the majority, these parts were developed and tested mostly independently, but of course also evaluated as a whole (see the following Sections 9.3 and 9.5). In this section, a brief overview is given for these components.

The topic clustering is based on the publication data in the annual document sets. By using LDA for the clustering, the output data are not mere document clusters. They are also instantiations of the topic model which is provided by the approach. The topic model is – in contrast to the respective document clusters – a conceptual representation of the topics. The term (and reference) probabilities for each topic are thus created as well. The clustering itself uses the features of the publications as described in Section 5.2.

The parameter setting also includes the testing and training of the similarity calculation between documents. The similarity between documents is calculated according to various features. One main part of this thesis was to determine the suitable features for this task. Again, the error rate and the threat of overfitting increase with a higher number of features (cf. p. 65). However, an approach with too few features for the similarity calculation runs the risk of underfitting or a superficial matching.

The similarities and thus the connections are calculated for clusters in different time periods, i.e. years. A topic that is found in the most recent year is compared to topics in the preceding years. A connection is built as soon as the threshold value t_c is exceeded. This connection represents the fact that a topic is continued over a time span of at least two years. The underlying assumption is that a topic cannot be new or innovative if it has a predecessor with such a high similarity. Thus, only those clusters that have no predecessor are selected for presentation to the end user.

The main part of the similarity calculation relies on the term probabilities of the topics. However, as the vocabulary changes over time for some topics, necessary connections might not be built. On the other hand, even after the removal of stopwords there might be terms that appear in various topics, maybe in different contexts or with different meanings, which suggest a factual non-existent similarity between topics. Because of that, the references are used as well. A weighting w_r for the references in

comparison to the terms had to be found. Furthermore, a threshold value t_c had to be selected to denote when two clusters were "connected". All these parameters are discussed, set and evaluated in Section 9.3.

In the third and final step, the emerging topic candidates are presented to the end user. These candidates are those documents that 1) belonged to unconnected topics and 2) deviated in their bibliometric feature values. The "deviation" had to be measured according to the usual values for the bibliometric features in established topics. For this, patterns in established and emerging topics are discussed in Section 9.4. The goal was to present the set of emerging topic candidates for intellectual inspection to an end user.

Each of these components can be developed and tested individually. For instance, the topic clustering can be developed independently and the results can be evaluated solely on the Precision and Recall of the topics. In contrast, the aggregated evaluation of all components is difficult, as will be discussed later in more details. In particular, the concept of a topic, its boundaries and its assignment to certain documents are all mere points of discussion.

In fact, in one project in which a system for topic classification was to be developed, the results of the Machine Learning algorithm were supposed to be evaluated with those of human experts. What had not been foreseen was the high disagreement between the three Computer Science experts and the classification in the dataset (see Table 25). In the end, the initially intended setting in which the experts' decision was compared to the assignment of the algorithm was nearly impossible.

Table 25: Agreement in class assignment for 100 documents by experts and classification system.

	Expert 1	**Expert 2**	**Expert 3**	**Classification system**
Expert 1	100%	61%	48%	39%
Expert 2		100%	46%	35%
Expert 3			100%	26%

Source: Own illustration

This experience showed the difficulty in assessing the quality of a classification system. Similarly to that project, in this thesis the codification of documents with regard to their topics' novelty had to be verified. Thus, one major task was to create a dataset for the automatic assessment of the system. Such a dataset facilitated on the one hand tasks like the parameter setting that required the repeated calculation of the evaluation metrics. On the other hand, the acquisition of an external validation source was meant to ensure credibility of the results. In the following section, the main datasets (Training and Test Sets) used in this thesis are discussed. Parts of them – namely of the NISTEP dataset – have already been introduced and applied (see Chapter 6 and 8). However, in the context of the parameter setting, rule derivation and evaluation of the proposed system, they are introduced again as a whole. Also, even though the initial dataset is the same, the selected parts differed for this chapter and Chapter 6.

9.2 Datasets

There were in total two bases for the datasets. One dataset was generated for the parameter setting of LDA in this thesis. It was established with conference proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) and the associated track information. The track information enabled the clustering of the respective documents for the Gold Standard. The additional bibliometric data was extracted from Scopus.

The second dataset was used for the derivation of the rules, the testing of the combination of the extended LDA and the rules and the evaluation of the overall approach. For these purposes, it was necessary to split the dataset in a Test Set and a Training Set in advance. The dataset was generated with the data from the NISTEP Science Map report 2008 (Saka, Igami and Kuwahara 2010). The information about the underlying disciplines was used to separate the data for the Test and Training Set. Overall, the dataset covered 10 out of 22 disciplines in the publication year 2007. The two datasets, one for the rules and the combination and one for the evaluation, consisted of different disciplines: 6 for training, 4 for testing.

The datasets are described in more detail in the following. The first dataset which was used for the parameter tuning in LDA is called “Training Set for LDA”. The other two datasets were labelled “Training Set for Rules and Overall Approach” and “Test Set for Overall Approach”. Their separation enabled a step by step training of the components and the final (unbiased) evaluation.

9.2.1 Training Set for LDA

As a Gold Standard for the clustering with LDA, documents were collected in Scopus that represented publications of the AAAI conference in the years between 2005 and 2010. For 2009, no documents could be found. The conference had approximately 15 tracks per year and 1,220 documents in total (see Table 26). The tracks were used to show one (of many) possible mapping that served as basis for the evaluation of the clustering approach.

Table 26: Overview of the Training Set.

Year	# tracks	# documents	Arbitrary tracks (#documents)
2005	19	224	1 (19)
2006	14	263	4 (83)
2007	15	261	4 (71)
2008	11	222	4 (67)
2010	15	250	6 (72)
total	74 (52)	1220	19 (312)

Source: Scopus, own calculations and illustration

Notes: Arbitrary tracks denote tracks that covered – by definition – documents from different topics, e.g. “umbrella topics”.

Even though there might be many possible clusterings (see Section 4.1.2 for Unsupervised Learning), it can be assumed that all documents in each track share at least one leitmotif or a common topic. Even

though some topics seemed arbitrary or ambiguous, e.g. “Intelligent Systems Demonstration”, they were included in the dataset to check the system’s ability to handle them. The number of such ambiguous tracks and the associated documents are listed in the last column in Table 26.

With the help of the so attained Gold Standard, the respective evaluation metrics for the LDA approach could be calculated (see Section 4.1.3). Without a sample of possible clusters, the evaluation would be restricted to non-quality evaluation metrics (e.g. equal distribution/cluster size) or a manual assessment in each application of the approach. With the calculation of Recall, Precision and F-Measure, a quantitative comparison of the results was enabled.

9.2.2 Training Set for Rules and Overall Approach

The Training Set for the rules and their combination with the LDA approach was equal to the dataset presented in Section 8.1. The origin, potential and limitations of this dataset were already explained in Sections 6.3.1 and 8.1. For the disciplines and respective document numbers see Table 18.

9.2.3 Test Set for Overall Approach

The Test Set was created in an analogous way as the Training Set. Thus, it was also based on the NISTEP dataset as described before (see Sections 6.3.1 and 8.1 in particular). However, it covered other disciplines in the year 2007 (see Table 27). The former disciplines and thus their associated methods were mapped on this dataset so that the approaches could be tested on a novel but similar and genuine dataset.

Table 27: Overview of distribution of document types in total numbers in the Test Set.

Discipline	New documents	Old documents	Total
Chemistry	53 (11%)	450 (89%)	503 (100%)
Mathematics	2 (5%)	38 (95%)	40 (100%)
Materials Science	18 (16%)	92 (84%)	110 (100%)
Space Science	10 (15%)	57 (85%)	67 (100%)
Total	83 (12%)	637 (88%)	720 (100%)

Source: WoS, own calculations and illustration

Notes: The shares in brackets are calculated in respect to each discipline.

For the evaluation, the approach uses only the information that would be available in a real case scenario, i.e. the textual information of the documents and the bibliometric data. This excludes the information whether a document is new. This particular information is however available for the later evaluation (and only for that), when the emerging topic candidates have been presented by the approach. Only then is the information about the new documents used to assess the quality of the candidate set. In particular, the share of new documents in the candidate set is calculated (= Precision, see Section 4.1.3). Furthermore, the number of new documents that is not contained in the dataset is derived (i.e. “losses” in new documents, share = 1 – Recall). In this way, it can be assessed how well the approach performs and whether it succeeds in separating the new documents from the rest or not.

9.3 Parameter Settings

9.3.1 LDA

The clustering results of the LDA approach for the varying parameter settings were compared to the above explained Gold Standard. As evaluation metrics, the adapted form of the Recall and the Precision as well as the $F_{0.5}$-Measure (as described in Section 4.1.3) were calculated. The clustering was performed on a yearly basis. For each testing of different parameters for LDA, the Recall and Precision values of the years were averaged to form one value. Therefore, the Recall and Precision were first calculated for each track and for each cluster respectively. For the overall value, the values for all clusters and tracks in all years were averaged. In addition, the derived value for the F-Measure is used to compare the different parameter settings. Parameters or parameter combinations are chosen according to the F-Measure values.

The aforementioned documents of arbitrary tracks might influence the Recall and Precision values as follows: When these documents are clustered separately instead of in combination with other documents, the Precision value goes up, because the other clusters are purer. Recall is unaffected as long as the documents are kept together in whatever form. Nevertheless, this might always lead to Recall and Precision values below 1.0 because of the distribution of these documents among other topics.[88]

For LDA, the following parameters were tested and adjusted:

- K,
- α,
- β,
- γ,
- threshold t_w,
- input data F_t
- weighting of the input data w_t

Because of the large number of different parameters, the parameters were tested consecutively. The sequence was:

1. α and n (K respectively)
2. β and γ
3. t_w
4. F_t (abstract, title, ...) and w_t

---

88 It can be shown that for an evenly distribution of M_t documents in an arbitrary topic t among K clusters, where each cluster k contains n_k documents in total, Precision can be $P = 1 - \sum_K \frac{M_t/K}{n_k}$ and Recall can be $R = 1 + \frac{1-K}{K*T}$ at maximum for T topics in total. Thus, the maximum value for Recall and Precision for the year 2005 are 1 and approximately 0.98 respectively; for the total dataset they account for approximately 1 (0.997) and approximately 0.99 respectively. This holds for K = 5, which is the lowest value tested in the following. However, with increasing values for K, the Precision value approaches 1.

All other parameters were set to initial values that were based on related work. The according reasoning is given in Chapter 5.

In the first parameter estimation, *n* and α were varied. α was either fixed to values between 0.1 and 1.0 or depended on the number of topics *K*. Usually, α is set to $\alpha = \frac{50}{K}$ (see Section 5.2 for discussion). But this works best if $K \geq 50$. Thus, both variance were tested – α with the fixed values between 0 and 1 and with the flexible value $\frac{50}{K}$. It follows from the latter, that the number of documents *M* and the (parameterized) number of documents per topic *n* determine not only the number of topics $K = \frac{M}{n}$, also *K* and thus *n* in turn influence α.

Table 28 shows the parameter settings that were tested in the first run. Fixed values are indicated by the value 1 in the last column, which denotes the number of variations of a parameter.

Table 28: Parameter settings for testing α and *n*.

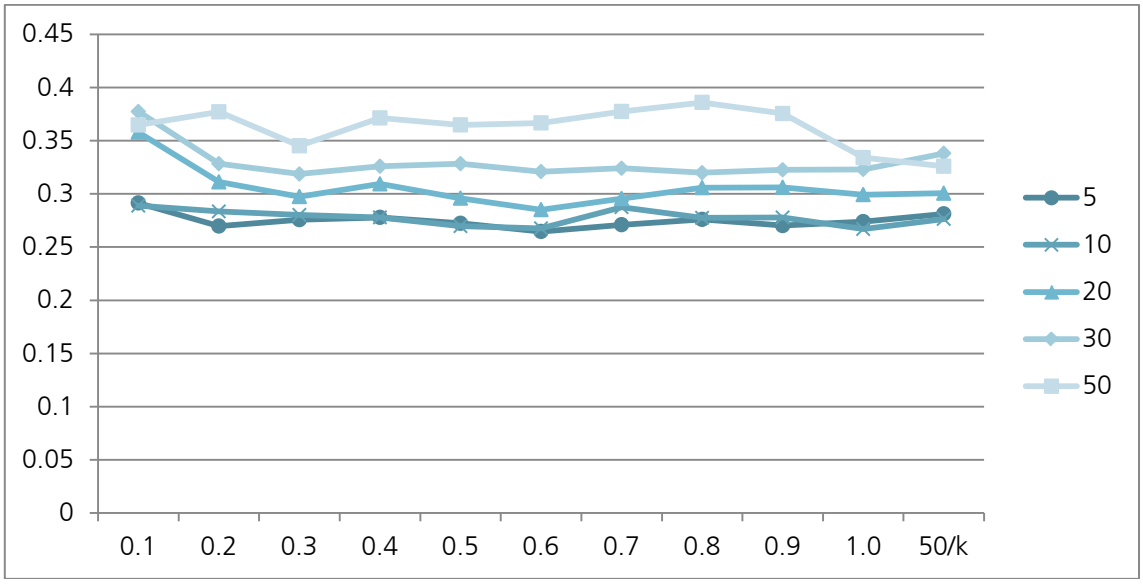
Parameter	**Fixed value(s)**	**Variable value**	**Number of different values**
α	0.1, 0.2, ...,1.0	50/*K*	11
K	---	*M*/*n*	1
n	5, 10, 20, 30, 50	---	5
β	0.01	---	1
Γ	0.01	---	1
F_t	abstract	---	1
w_t	1	---	1
t_w	0.5	---	1

Source: Own illustration

Therefore, α could take any value from 0.1 to 1.0 in 0.1 steps or $\frac{50}{K}$ and *n* was set to values 5, 10, 20, 30 and 50 (see Table 28). Since many of the values of *n* resulted in *K*<50, α had in most cases values above 1.0 for its variable setting. All other values were – for the time being – fixed to their default values, i.e. β = γ = 0.01 and t_w=0.5.

Figure 31 shows the F-Measure values for these test runs. The colour of the lines represents the value of *n* while the x-axis shows the varying values for α. The best parameter combination was α=0.8, *n*=50 with a Recall and Precision value of approx. 0.50 and 0.31 respectively. Since Precision prefers larger clusters, here the best results overall were achieved with *n* = 5, whereas the best results for Recall are achieved with large values for *n*. On the other hand, the Recall was slightly improved with a small value for α, since this seemed to make the topic representation more distinct. For Precision, no tendency for any specific value of α could be seen – the variance for the different parameter settings was 0.0044.

Figure 31: The F-Measure for different values of *n* (colour of lines) and α (x-axis).



Source: Scopus, own calculations and illustration

Next, different values for β and γ were tested ranging from 0.01 to 0.09 and from 0.1 to 1.0 in 0.01 or 0.1 steps respectively.

Table 29 lists the top 10 values for F-Measure, Recall and Precision and the respective values of β and γ. Recall varies between 0.55 and 0.57, Precision between 0.31 and 0.33 and the F-Measure between 0.39 and 0.41. The high value for β suggests that a sparse word distribution is inappropriate. Rather, a high ambiguity of terms was favoured, i.e. a scenario where each term belonged to a multiple set of topics. In contrast to that, rather low values for γ were preferred in terms of F-Measure and especially in terms of Recall.

Table 29: Top 10 values for Recall, Precision and F-Measure for varying values of β and γ.

Rank	F-Measure	β	γ	Rank	Recall	β	γ	Rank	Precision	β	γ
1	0.41	1.0	0.08	1	0.57	1.0	0.4	1	0.33	1.0	0.08
2	0.40	1.0	0.6	2	0.56	1.0	0.04	2	0.32	0.03	0.9
3	0.40	0.6	0.1	3	0.56	1.0	0.05	3	0.32	0.06	0.06
4	0.40	1.0	0.01	4	0.56	0.8	0.9	4	0.32	0.05	0.3
5	0.40	0.06	0.06	5	0.55	1.0	0.01	5	0.32	0.9	0.6
6	0.40	1.0	0.05	6	0.55	0.8	0.05	6	0.31	1.0	0.6
7	0.40	0.9	1.0	7	0.55	1.0	0.02	7	0.31	0.9	1.0
8	0.40	1.0	0.7	8	0.55	1.0	0.03	8	0.31	0.2	0.08
9	0.39	0.9	0.6	9	0.55	1.0	0.3	9	0.31	0.2	0.02
10	0.39	0.2	0.2	10	0.55	1.0	0.06	10	0.31	0.6	0.1

Source: Scopus, own calculations and illustration

In all three measures, β = 1.0 seemed to be the best choice. The combination of β = 1.0 and γ = 0.08 ranks first in terms of Precision and F-Measure and is ranked on position 11 with a value of 0.55 for Recall. Thus, β was fixed to 1.0 and γ to 0.08.

Then, different values for the term occurrence threshold t_w were tested (Table 30). Values between 0.1 and 1.0 were used in this analysis to determine again the best setting for Recall, Precision and in particular the F-Measure.

Table 30: Values for Recall, Precision and F-Measure for varying values of threshold t_w.

	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Recall	0.65	0.65	0.66	0.68	0.65	0.67	0.68	0.66	0.66	0.64
Precision	0.32	0.31	0.30	0.30	0.35	0.32	0.33	0.31	0.32	0.35
F-Measure	0.43	0.42	0.42	0.42	0.46	0.44	0.44	0.42	0.43	0.45

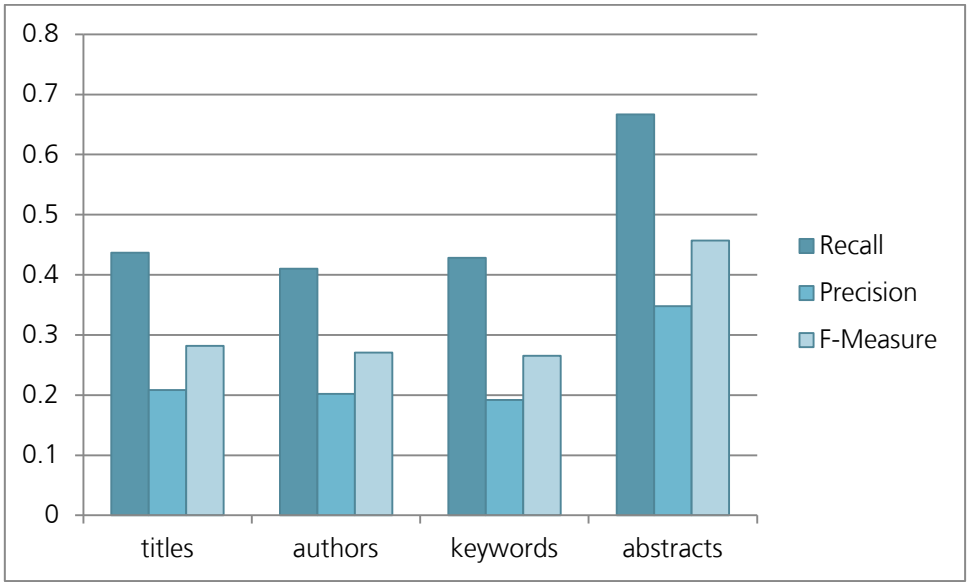
Source: Scopus, own calculations and illustration

Best values for Recall were achieved with t_w = 0.7, while Precision and F-Measure were highest for t_w = 0.5. Thus, t_w was fixed to 0.5 in the following. Recall now accounts for approximately two thirds, i.e. for each topic there exists a cluster that covers on average at least two thirds of the respective documents.

First, different input texts F_t, i.e. abstracts, keywords, titles and authors, and then different combinations of the aforementioned were tested to define the best mixture of inputs. When combining them, the inputs were weighted differently to account for a varying importance of terms from the respective fields. However, a distinctive topic modelling like for terms and references is not made for the different fields. Thus, after the weighting, all terms are used in the same term topic distribution.

Alternatives for the abstract as input were tested, but Recall, Precision and F-Measure values only worsened in comparison (Figure 32). This may stem from the fact that the abstract was used as a fixed input so far and the remaining parameters were adjusted respectively. Alternatively it may indeed be the best indicator for topical relatedness.

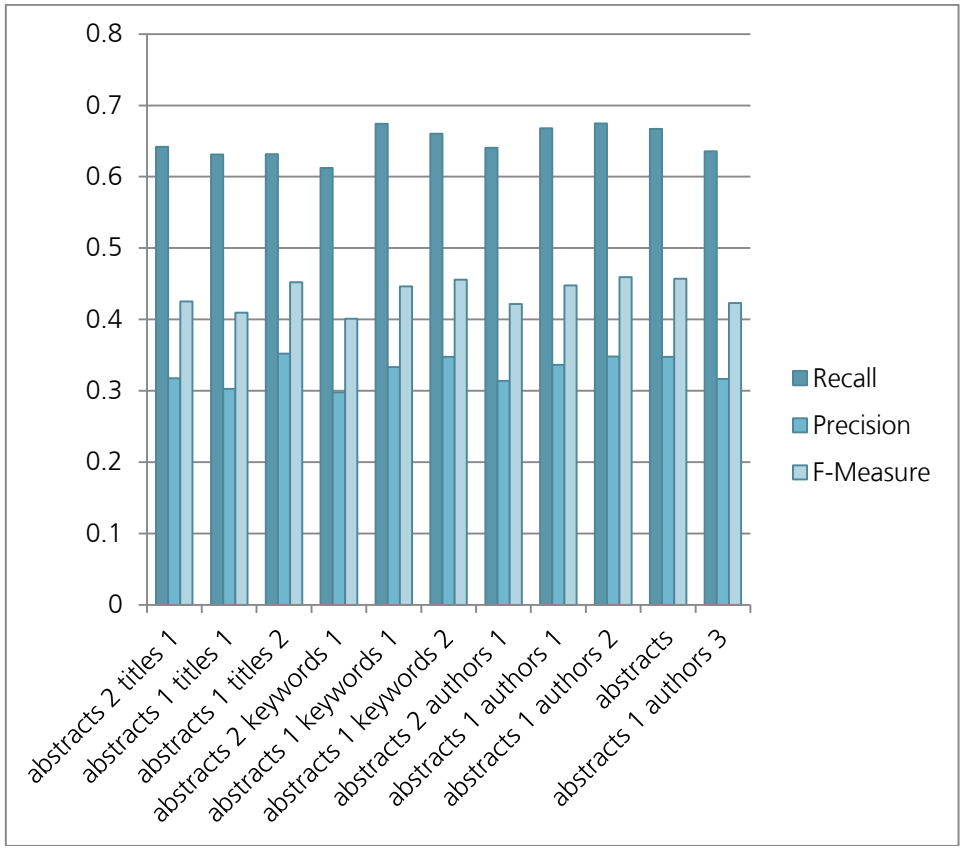
Figure 32: Recall, Precision and F-Measure values for different textual inputs F_t.



Source: Scopus, own calculations and illustration

The combination of abstracts with titles, keywords and authors with different weightings follows. Since abstracts performed best before, the usage of abstracts with the other text fields was combined. Figure 33 shows the results.

Figure 33: Recall, Precision and F-Measure for different combinations of abstracts and other textual input features.



Source: Scopus, own calculations and illustration

There are few combinations of inputs that lead to better results than the sole usage of abstracts; Recall and Precision measures can be improved by an equally weighted combination of abstracts with authors or keywords. However, the best results in terms of F-Measure are achieved with abstracts and authors with a ratio in weighting of 1:2.

Further testing with the combination of three input files, e.g. abstracts, authors and keywords did not improve the results. Thus, abstracts and authors are used as input data with a weighting of 1:2. Hence, the parameter setting ends with the values shown in Table 31.

Table 31: Final parameter settings for LDA.

Parameter	Setting
α	0.8
K	M/n
n	50
β	1.0
⊤	0.08
F_t	Abstracts, authors
w_t	1, 2
t_w	0.5

Source: Own calculations and illustration

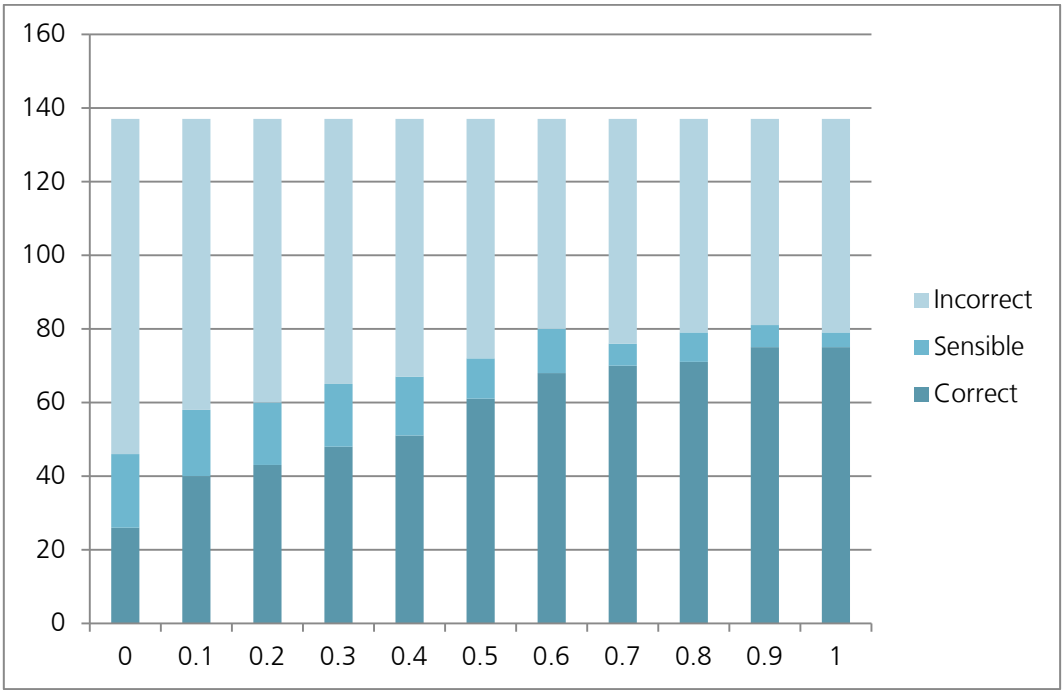
9.3.2 Similarity between Topics

For the similarity between topics, the parameters concerned the weighting of the reference in comparison to the term distribution w_r, the threshold for a connection between two topics t_c and the number of years Y considered for the similarity calculation. Table 6 on p. 87 lists a complete set of parameters that were adjusted for the similarity calculation between topics.

The threshold t_c was first set to 0.0 so that all matches (with all similarities) were covered in the result set. The values for w_r were varied between 0.0 and 1.0 in 0.1-steps. Again, the set of conference tracks of the AAAI conference were used as a Gold Standard. This time, the tracks were used as input for already built clusters to keep the parameter estimation free of errors due to a former imprecise clustering. The connections between the tracks were assessed manually. Therefore, the labels of the respective tracks were compared. The number of connections that were correct was calculated for each value of w_r first. Connections between topics that did not necessarily represent the same concept but which where related were weighted 0.5 (in contrast to a correct connection with the value of 1.0).

The number of correct (or sensible) connections was summed up over all years. No normalization could be performed because there was no standard for the exact number of connections that should be established. However, the more correct connections that could be confirmed manually the better.

Figure 34: Number of incorrect and correct connections for varying values of w_r.

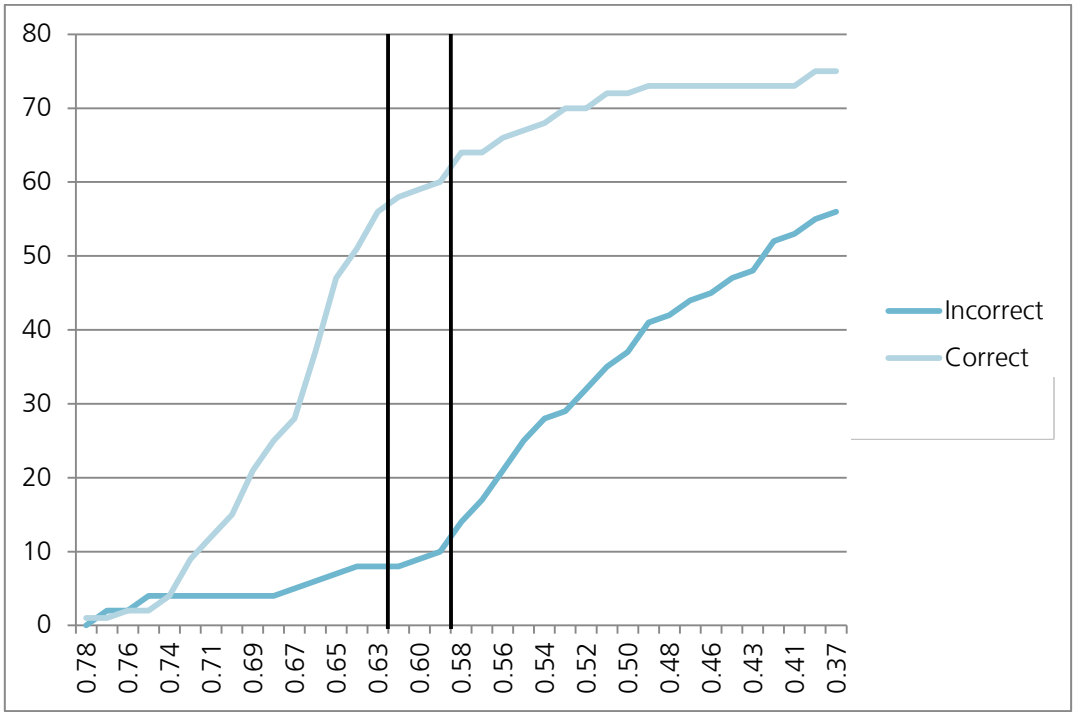


Source: Scopus, own calculations and illustration

Figure 34 shows the number of connections that were correct, incorrect or at least sensible. The minimum number of incorrect connections is achieved with a weighting of 0.9 of the references. With w_r = 0.9 and w_r = 1.0, the highest number of correct connections is accomplished (75 out of 137). Thus, the approach was run with w_r = 0.9 in the following. While the parameters in the LDA approach did not indicate the weighting between the importance of the references and the terms used in the approach, the parameter w_r for the connections can be seen as a clear sign for the importance of references for the connection of topics. A weight of 0.9 for the references (which reduces the relative weight of the terms to 0.1) shows that the unstable vocabulary in topics can lead to false connections. Conversely, a connection via references is more reliable.

Figure 35 shows the absolute number of incorrect and correct connections for a diminishing threshold value t_c from left to right. As can be expected, the number of incorrect connections increases with a diminishing threshold, i.e. with a lower threshold more and more incorrect connections are accepted. However, at the same time correct connections are added as not all of them are covered by the highest possible threshold values. A steep increase can be observed for threshold values $t_c > 0.62$ (see vertical line on the left hand side of Figure 35). After that, the increase in correct connections more or less levels out. Conversely, up until that point the number of incorrect connections is relatively low (8 incorrect to 58 correct connections). However, while the line for correct connections flattens for $t_c > 0.62$, the one for incorrect connections steepens. For a value of $t_c < 0.58$ the observation is reversed (see vertical line on the right hand side of Figure 35). Thus, the approach was applied with a threshold value of t_c = 0.58.

Figure 35: Absolute number of correct and incorrect connections for varying threshold values t_c.



Source: Scopus, own calculations and illustration

Since connections were built with all topics in all previous years, the question was whether connections with topics in a shorter period might not be more reliable than others. A comparison of the connected topics in the different time periods confirmed that the set of topics that is connected in one year is independent of the target year for the connections. Thus, it is better to restrain the connections to those with the topics in the previous year, since there is no information lost for the detection of unconnected topics.[89] However, the distinction between correct and incorrect connections showed that the connections with a time window of one year are more accurate.

9.3.3 Overall Approach

Table 32 shows the final parameters for the connections. With these parameters and the above determined parameters for LDA, the approach was run again on the whole dataset. After restricting the connections on consecutive years, the final year of the evaluation was set to 2008.

Table 32: The final setting of the parameters for the connection establishment.

Parameter	Description
t_c	0.58
w_r	0.9
Y	1 year

Source: Own illustration

89 Except for the information of the year 2010 for which there was no foregoing year in the Training Set.

Table 33: Clusters generated by the approach with the so far set parameters (number of citations are derived from the Scopus database).

ClusterID	Document Title	Track	Cit
1	Manifold integration with markov random walks	Knowledge Representation, Logic, and Information Systems	0
3	Voting on multiattribute domains with cyclic preferential dependencies	Constraints, Satisfiability, and Search	4
	Determining possible and necessary winners under common voting rules given partial orders	Constraints, Satisfiability, and Search	6
	Minimal contraction of preference relations	Knowledge Representation, Logic, and Information Systems	0
6	Non-monotonic temporal logics that facilitate elaboration tolerant revision of goals	Knowledge Representation, Logic, and Information Systems	0
14	On the dimensionality of voting games	Agents, Game Theory, Auctions, and Mechanism Design	1
	On range of skill	Constraints, Satisfiability, and Search	0
	Manipulating the quota in weighted voting games	Constraints, Satisfiability, and Search	1
19	Extending the knowledge compilation map: Krom, horn, affine and beyond	Knowledge Representation, Logic, and Information Systems	0
	Efficient haplotype inference with Answer Set Programming	Knowledge Representation, Logic, and Information Systems	0
	Parallel belief revision	Knowledge Representation, Logic, and Information Systems	2
	A reductive semantics for counting and choice in answer set programming	Knowledge Representation, Logic, and Information Systems	9
	Efficient haplotype inference with answer set programming	Knowledge Representation, Logic, and Information Systems	3
	A meta-programming technique for debugging answer-set programs	Knowledge Representation, Logic, and Information Systems	0
	A semantic approach for iterated revision in possibilistic logic	Knowledge Representation, Logic, and Information Systems	0
	Yoopick: A combinatorial sports prediction market	Intelligent Systems Demonstrations	1
	Horn complements: Towards Horn-to-Horn belief revision	Knowledge Representation, Logic, and Information Systems	1
36	A scalable jointree algorithm for diagnosability	Knowledge Representation, Logic, and Information Systems	1
37	Prime implicate normal form for ALC concepts	Knowledge Representation, Logic, and Information Systems	1

Source: Scopus, own calculations and illustration

7 clusters were selected by the approach as not having any connections with previous research (Table 33). The majority (namely clusters 1, 6, 36 and 37) contained only one document. For the other clusters, the majority of the documents in a cluster originated from the same track, i.e. 2/3 of the documents of clusters 3 and 14 and 8/9 of cluster 19 had been assigned to one track at the conference and thus showed a high topical relatedness. With regard to the innovativeness, the citation rate in the first three years after publication does not provide many hints – as already discussed in Chapter 8 this comes to no surprise. However, some documents in clusters 3 and 19 receive extraordinarily high citation rates. The associated topics are "Voting rules and preference relations" and "Efficient answer set programming".

A genuine assessment of a cluster's innovativeness and novelty would demand the judgement by an expert. Thus, the assessment of the topics was restricted to a description which gives the reader the opportunity to wager for himself whether he would deem the topic as emerging in the year 2008 or not. The topics of the respective clusters are:

- Cluster 1: The paper presents an approach for manifold integration as an alternative to RKKS and DISTATIS. According to Google Scholar, the paper so far[90] has been cited by 6 publications, of which all but one stem from at least one of the authors of the original paper. Interestingly enough, this particular publication deals with answer set programming like Cluster 19.
- Cluster 6: The authors show a new form of non-monotonic logics that allow revision and exception handling of goals. The paper aims at directives given to robotic agents that need to be adjusted to current situations or need to be revised because of new ideas, thoughts or requirements of the human handler. The authors express the necessity for such options: "In rescue and recovery situations with robots being directed by humans, there is often so much chaos together with the gradual trickling of information and misinformation that the human supervisors may have to revise their directives to the robots quite often" (Baral and Zhao 2008, p. 406).
- Cluster 14: The cluster is concerned with the assessment of voting games. The first two papers listed in Table 33 deal with dimensionality and running time estimations of games. The last paper estimates the power of individual players in a game depending on their quota, or more precisely, it calculates whether a player is reduced to a dummy role given his and other players' quota.
- Cluster 19: The majority of the papers in this cluster deal with answer set programming. Other papers were added that are concerned with probability related prediction, beliefs or horn clauses.
- Cluster 36: This paper offers a new approach for the diagnosibility problem that avoids typical issues of the so far applied twin plant method. According to Google Scholar, it has been cited at least 4 times by other authors.
- Cluster 37: A new normal form for expressions in ALC is provided and compared to former normal forms. In particular the fact whether a concept subsumes another concept can be more easily and efficiently verified with the new normal form. 9 citations (including 1 selfcitation) can be found in Google Scholar for this paper.

90 Status February 2013.

9.4 Rules

For the second step of the approach, rules to identify outliers in the document sets had to be found. These rules were derived from “typical” feature values for publications in a discipline. The distinction in the NISTEP dataset between “new” and “old” topics helped to specifically identify those features for which “new” publications deviate.

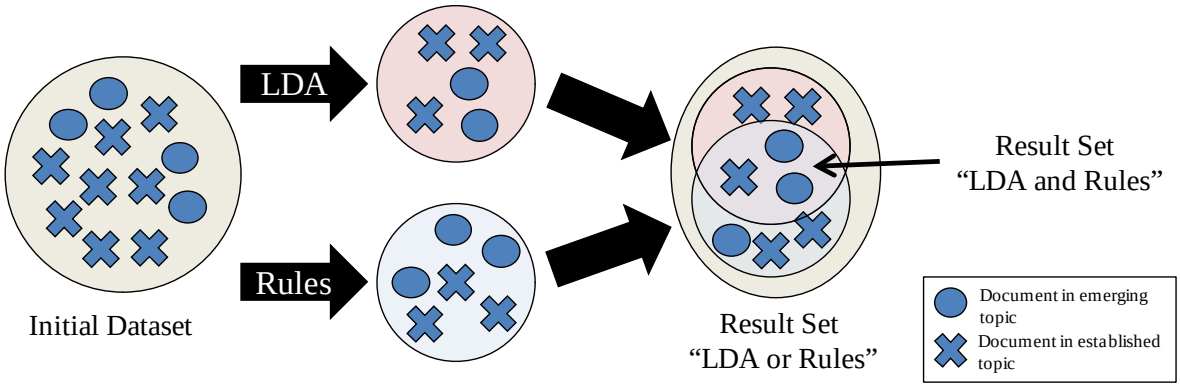
9.4.1 Methodology

For each discipline, a description of notable deviations is shown with the help of diagrams. Outstanding distributions of new and old topics with regard to the features are discussed. Furthermore rules for the detection of emerging topics are derived with a Machine Learning approach. The results are compared for both methods to find a set of sound rules for the emerging topic selection. This set builds the foundation for the second part of the approach, the selection of single documents as emerging topic candidates.

For the final feature selection, the Information Gain (cf. Section 4.1.3) was calculated to deduce Conjunctive Rules. Conjunctive Rules are a set of rules as small as possible that enables a classification in the dataset (Witten and Frank 2005, pp. 408f). Thus, they use one or more features to make an implication for the class value. They facilitate the selection of features and feature values for the emerging topic detection.

The approximation of a uniform distribution was realized by Resampling in Weka (see p.72, Witten and Frank 2005, p. 400). The size of the new sample was for each discipline equal to the initial size. The resampled sets were then used to derive Conjunctive Rules.

Figure 36: The combination of both approaches to get a more specific or wider result set.



Source: Own illustration

Both approaches, the rules derived from the Machine Learning approach as well as the initial LDA approach can be applied solely as well as in combination. Figure 36 shows exemplarily the correlation between the respective result sets. The initial dataset can be used as input for either LDA or the rules, which might result in different or similar document sets. The union of both sets is the or-conjunction, i.e. either LDA or the rules would label the respective documents as new documents. In contrast to that, both approaches need to agree on the labelling for new documents for the result set in their and-combination, which equals their overlap.

9.4.2 Descriptive Analysis and Conjunctive Rules

For each document, the values for the features in Section 6.2 were calculated. The individual values were also aggregated to determine the discipline average. The documents were ordered and grouped according to their feature values. For instance, all documents that shared the value for the journal size were summarized. Thanks to the determined class values, the share of new documents for a specific feature value or even a range thereof can be calculated.

In order to evaluate the rules, a restriction to a maximum value as well as a minimum value was tested. In this way, also rules that define a range for a feature value can be visualized. Therefore, two thresholds are introduced that help monitor the data for increasing or decreasing values:

- A lower threshold measures the number of documents (new and old) that lie above a certain value.
- An upper threshold quantifies all documents that have feature values below a specific value.

The upper threshold simulates a rule that cuts away all documents that lie above a specific value. The share of new documents in the so gained document set can be calculated to indicate the ratio of new and old documents in the remaining set. This simulates the effect of a rule that uses the threshold to separate new and old documents; the share of new documents in the remainder set indicates how well such a rule performs on the Training Set. A high share of new documents, i.e. 100% in the best case when only new documents are left, signifies a well applicable threshold value. Since this facilitates the interpretation of the rules, these shares of new documents are shown in the following.

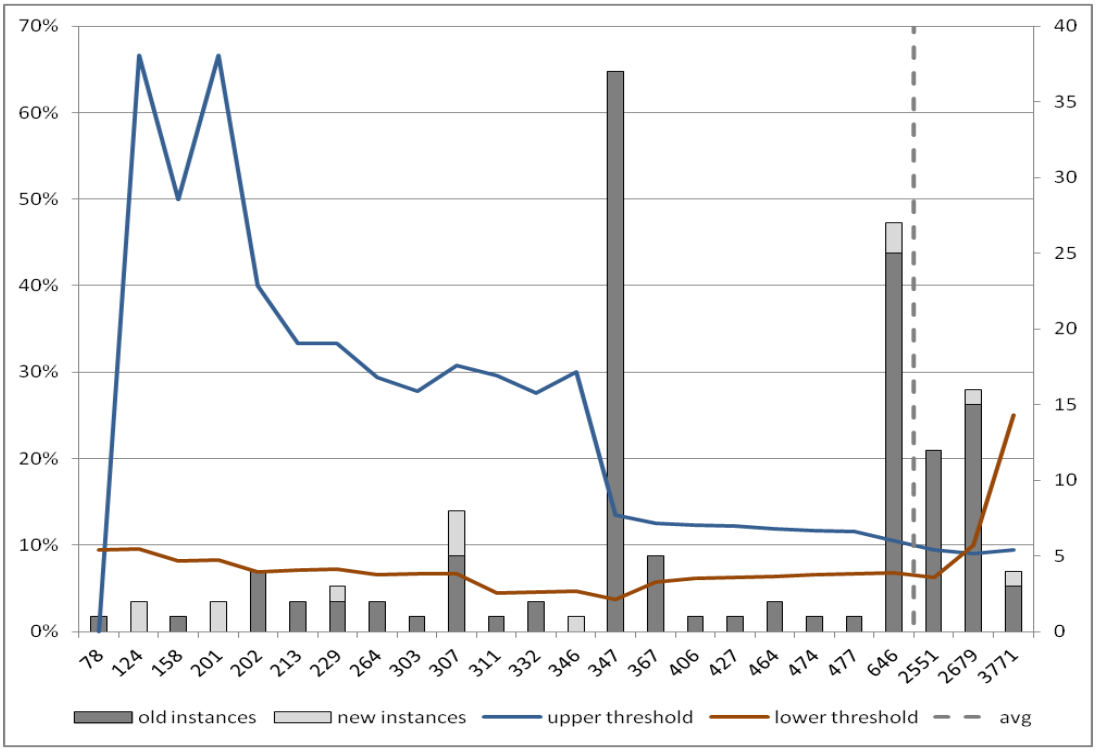
Figure 37 is one exemplary illustration for the following analysis and is used for explaining the interpretation of the respective graphs. It shows the share of new topics in a dataset for the upper and lower threshold. These threshold values are the basis for the rules that are derived later; each rule uses an upper and/or lower threshold to diminish the set of documents to a set in which all documents have values above and/or below the threshold values. Thus, measuring the share of new documents in a set of documents which have feature values above/below a certain value shows the usefulness of a (lower/upper) threshold of the same value. Therefore, the lines in Figure 37 show the share of new documents in the dataset for the thresholds. They enable the estimation of the quality of the result set for diminishing or increasing threshold values at a glance.

On the x-axis, the values for the feature (journal size in the category Molecular Biology & Genetics) are shown. The bars above illustrate how many of the documents were published in a journal of that size (y-axis on the right hand side). For other features, which include fractional values, the rounded values are shown for better readability and interpretation. The light grey bar denotes the absolute number of new instances, the dark grey bar that of the old instances.

The blue and the red line represent the shares for the respective threshold values as described above. The threshold value always corresponds to the value on the x-axis. Therefore, the threshold values increase from left to right. For the lower threshold, this leads to a narrower delineation and thus a decreasing size of the result set. The share of new documents in the resulting set is depicted by the lines (y-axis on the left hand side). For instance, the lowest feature value in the overall document set in terms of journal size is 78. The bar in Figure 37 shows that only one (old) document has this value (y-axis on the right hand side). If the upper threshold is set to this value, all other documents are excluded

from the result set. Therefore, the blue line representing the share of new documents in the result set shows a value of 0%. The lower threshold in turn represents the rule covering all documents with a feature value of at least 78. For the lowest value in a dataset this corresponds to no restriction. Thus, resulting and original dataset are equal. The share of new documents (approx. 9%) reflects the figures shown in Table 18. Conversely, for a value of 3771, there are 4 documents left of which only one is a new document. Thus, the lower threshold accounts in this case for 25%. For the value of the upper threshold, the initial set is covered completely for an upper threshold of 3771. Thus again, a value of approximately 9% can be observed.

Figure 37: Illustration of thresholds used in the descriptive analysis (journal size in Molecular Biology & Genetics).



Source: WoS, own calculations and illustrations

The immense number of publications with a journal size of 347 is excluded for an upper threshold of 346 or below. Since these cover only old documents, there is a steep increase in the percentage of new documents in the result set and thus in the blue line. However, seemingly both threshold values are sensitive for small document sets.

The grey dotted line in the graph denotes the average value in the respective discipline. It allows interpretations that rely on the comparison to the total discipline (in that specific year). In particular, it enables the deduction of rules regarding the average. For example: Documents with a feature value lower than average have a higher chance to be new etc. In this particular case, this additional view shows that by tendency, a lower threshold below the average is less useful than one above the average. Please note that in this case as well as in all other graphs, for better readability the average is always depicted as a line exactly in the middle between the two values between which it lies. However, the actual value might be closer to one of the two values between which it is illustrated.

In the following, the graphs for the different disciplines are shown. For each discipline the course of the upper and the lower threshold are depicted. The line denotes the respective share of new documents in the total document set with this threshold. Thus, the higher the share, the higher is the Precision of the feature/threshold combination and also the probability for a new document when selecting one document in the remaining document set randomly. Therefore, the threshold helps to sort out the old documents, but if it also loses many of the new documents it would be counterproductive.

For the deduction of rules, Conjunctive Rules were applied. Conjunctive Rules always select the best fitting features to separate the data according to the given class distribution. However, in order to get as many rules as possible, the attributes from the first Conjunctive Rule were exclude before applying the Conjunctive Rule a second (and a third, and a forth, …) time. This step was repeated as long as the Conjunctive Rule resulted in a class separation better than average (or better than the one given in the initial dataset). Each table presented for the rules consists of the following columns:

- Feature & Condition: This is the rule that was derived by the Machine Learning approach. The feature, feature value(s) and the relation (e.g. larger/smaller than, equal to or between) are given. Some rules consisted of multiple parts which were connected with an “and”-connection. In this case, the column also contains “AND” and the other conditions.
- Separated class: The rules either aimed at selecting a set of new or a set of old documents (cf. explanations for skimming and cutting given in Section 2.2). This column shows, whether the main part in the result set was new or old documents.
- New documents/Old documents: The last columns show the distribution of new and old documents in the set of selected documents and in the remainder set. The latter numbers are given in brackets to indicate that they are included in the part of the set that is “cut away”. However, these numbers are still of value to show how many false negative classifications (see Section 4.1.3) are introduced by that rule. However, more important in the context of this thesis are the respective values for the selected set. With their help, Recall and Precision (see Section 4.1.3) can be calculated for the result set (in the Training Set).

After the derivation of Conjunctive Rules, the rules were tested on the initial dataset. They were tested solely and in combination. The performance was measured with the $F_{0.5}$-Measure.

In a first testing, it became apparent that the “or”-combination of the tested rules was always inferior to any “and”-combination. Thus, in the following only the testing of “and”-combinations are described.

First, the “and”-conjunction for two and for all rules was tested. If the set of two rules had higher scores, it was extended with further rules. This step was repeated until no improvement could be achieved with further extensions.

If the best set in the initial setup was the combination of all available rules, the removal of single rules was tested instead. Again, this was performed until the $F_{0.5}$-Measure was equal or less to previous results.

Computer Science

The descriptive analysis shows that the new documents in Computer Science were published only in journals with an age between 3 and 27 (Figure 38). Thus, in the oldest or youngest journals in the set, only old documents were published.[91]

The most obvious reason behind this observation might be the already narrowed down topical coverage of the journals and the usage of proceedings in Computer Science. The topics of the older journals are partly very specific. One describes itself as focusing "on all telecommunications including telephone, telegraphy, facsimile, and point-to-point television, by electromagnetic propagation, including radio; wire; aerial, underground, coaxial, and submarine cables […]".[92] Current hot topics can be expected seldom in such journals. An outdated topical focus in established journals might repulse publications in emerging topics. However, as long as the readership is sufficiently large, there might be no need for a change (observable). On a similar notion, new developments may be first published at conferences where they might be more on the cutting edge and have a smaller time lag between submission and publication (Eckmann, Rocha and Wainer 2012). The first publication might make a (re-) publication in (older) journals obsolete.

91 Note that in the regression model (Chapter 6), no direct relation could be found for the respective aggregated discipline.

92 http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=26, last accessed 2013/04/24.

Figure 38: Journal age (Computer Science).

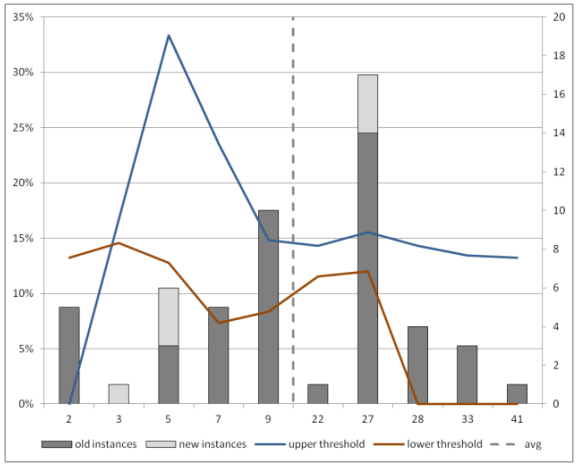


Figure 39: Age of the references (Computer Science).

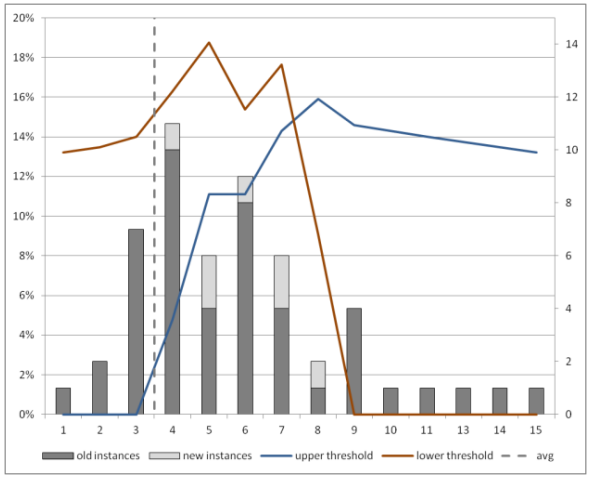


Figure 40: Fields of the journal (Computer Science).

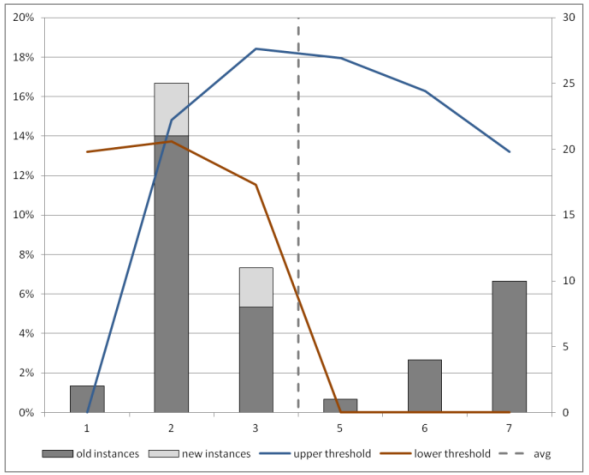


Figure 41: Fields of the references (Computer Science).

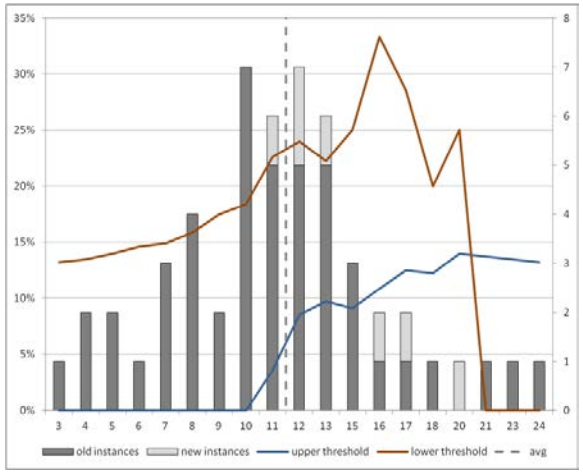


Figure 42: JIF (Computer Science).

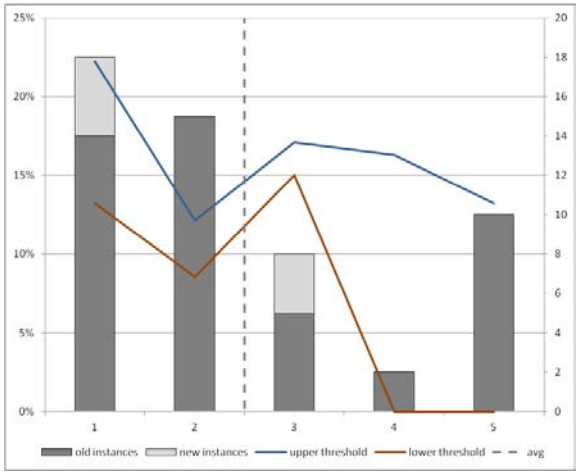
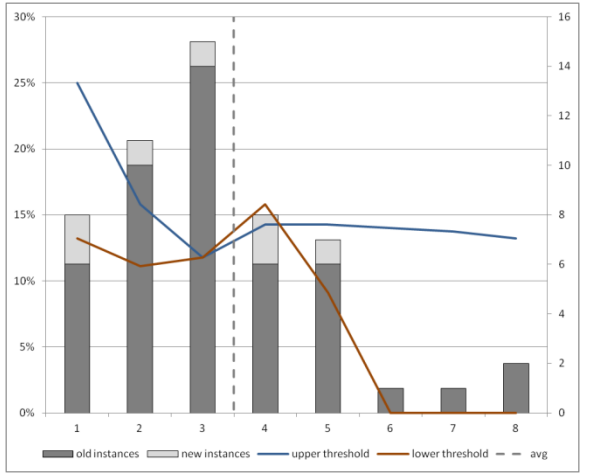


Figure 43: Number of authors (Computer Science).



Source: WoS, own calculations and illustrations

Furthermore, new documents can only be found in the set of documents with a reference age between 4 and 8 years (Figure 39). Also, the distribution clearly shows that new documents are restricted to those journals with only 2 or 3 fields or in other words below average (Figure 40). Contrary to that, most new documents' references cover more scientific fields than those of the average publication (Figure 41). Only the latter could be confirmed by the regression for the aggregated field (cf. Chapter 6). It is difficult to decide which of these would be the better measure for interdisciplinarity. However, the results suggest that the new documents are published in specialized journals but have to rely on foundations from different fields. The JIF and the number of authors also show a high tendency for lower values (Figure 42**,** Figure 43).

Table 34: Proposed rules for Computer Science.

Feature & condition	**Separated Class**	**New documents**		**Old documents**	
		Classified as new	**Classified as old**	**Classified as new**	**Classified as old**
Fields of Journals >4	Old	(25)	0	(20)	8
Age of References > 8	Old	(20)	5	(17)	11
Fields of References < 11	Old	(22)	3	(14)	14

Source: WoS, own calculations and illustrations

Table 34 shows the number of new and old documents in the set resulting from the application of the derived rules. As explained above, the first column lists the rule that is applied while the second column shows the target class of the rule. In this case, all rules are used to extract old documents, which can then be excluded to form a more precise document set. The following columns show how many of the new and old documents are covered by this rule. For instance, the third and fourth column show how many of the new documents are classified as new or old when the rule is applied. If a rule targets old documents, the documents that need to be inspected are those labelled old. The other documents are only implicitly tagged as new. Rather, they are in the resulting set which is due to further processing anyway and are therefore denoted in brackets. In other words: The table shows the number of true positive, true negative, false positive and false negative instances of a rule. However, whether "positive" is associated with new or old documents depends on the target of the rule, i.e. the separated class. Therefore, the table shows the number of new/old documents classified as new/old. Note in this context that the information whether a document is new or old is only known in the Training Set scenario and only for the evaluation.

The very first rule derived for the set of Computer Science publications is that by trend a number of journal fields higher than the average indicates a publication in an old topic (Table 34). Interestingly, 29% of the old documents can be filtered out by applying this rule. Therefore as already noted in the descriptive analysis, a high number of fields rather indicate an established topic.

The second rule shows that all publications with higher average age of references were old documents. However, a small loss in genuine new documents can be perceived for this rule in the resampled set.

As was observable from the descriptive analysis, the age of the references indeed proofed to be a well suited indicator.

The final rule found by the Machine Learning approach for Computer Science is even the most selective. The rule states that documents with fewer fields in the references than the average are by trend old documents. It describes that no or only few interdisciplinarity induces a smaller chance for new documents. In fact, 43% of the documents in the resampled set have a value of less than 11 fields and could be sorted out. In the genuine dataset, these accounted for 48%.

The rules thus tested on the initial dataset were:[93]

- R1: Fields of the journal < Average
- R2: Age of the references between 4 and 8
- R3: Fields of the references ≥ Average
- R4: Age of the journal ≤ 27
- R5: JIF < 4
- R6: Authors ≤ 5

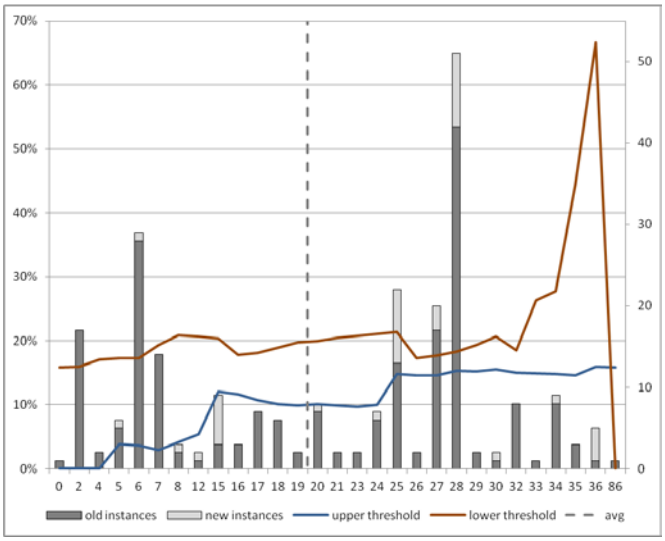
All the above rules were confirmed by the descriptive analysis. Rule 2, which corresponds to the second rule found with the conjunctive rules, was extended with the findings of the descriptive analysis to also exclude those documents with the youngest average age of their references.

By applying these rules, the $F_{0.5}$-Measure in Computer Science could be improved from 19% to 32%. The Precision increased to 43% while only one of the 7 new documents was sorted out by the rules. Rules 5 and 6 could be removed without changing the result set. A smaller set of rules is preferred for better appliance on other datasets (cf. definition of overfitting on p. 65). Removing R1 showed that this rule and R3 are complementary – a notion that was not perceivable from the descriptive analysis. Thus, the final rule set for Computer Science encompasses rules 1 to 4.

93 The rule country = 1, which was tested because of the results in the descriptive analysis, only led to worse results as a stand-alone rule as well as in combination with others.

Engineering

Figure 44: Age of the journal (Engineering).

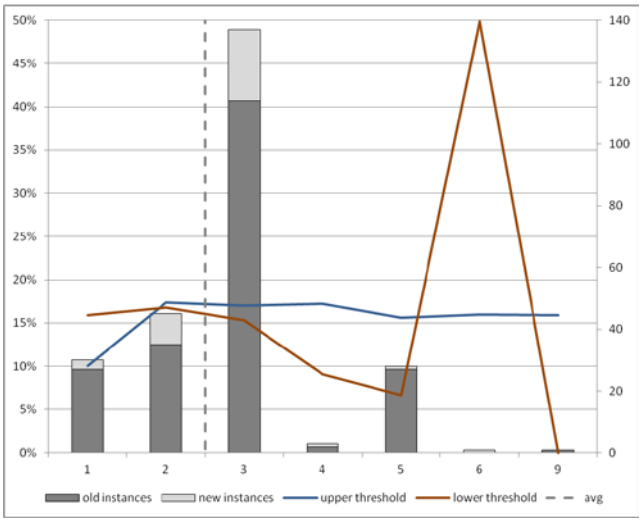


Source: WoS, own calculations and illustrations

For Engineering the slope for the journal age shows that the youngest journals do not contain any (since 2003) or only few (since 1995) new documents (Figure 44). The upper threshold suggests an increasing chance for new documents in older journals. Only the oldest journal with one document in total makes an exception. The regression in Chapter 6 corroborates these findings, as a positive significance for the journal age could be found.

Given the data, the introduction of a rule covering the "minimum age of a journal" seems advisable. In the rules tested below, the split is made for documents with a journal age of at least 8 years.

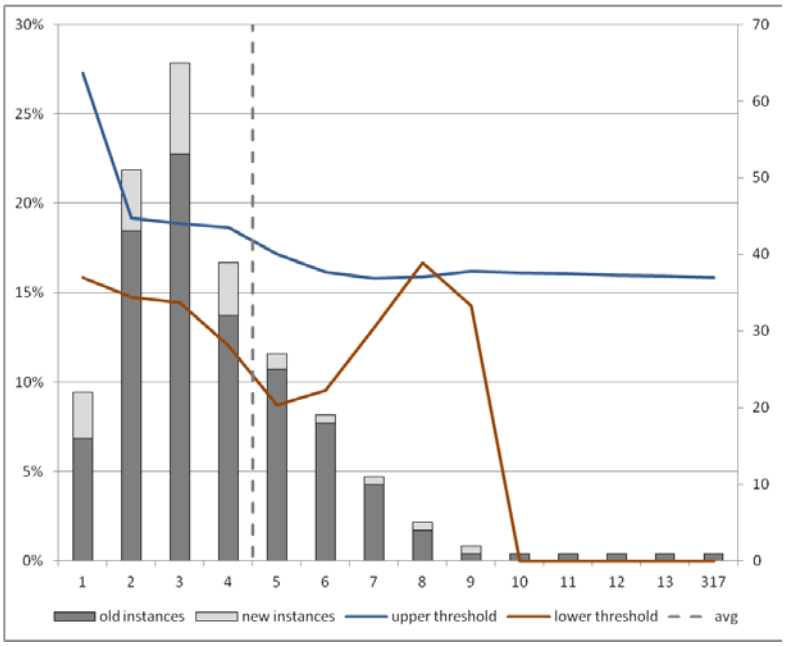
Figure 45: JIF (Engineering).



Source: WoS, own calculations and illustrations

Furthermore, most new documents were published in journals with a JIF below 4 or respectively around the discipline's average (Figure 45). Also, there were no new documents with more than 9 authors (Figure 46). The latter contradicts the findings in the regression model for Engineering (Chapter 6), which also bore a low significance.

Figure 46: Number of authors (Engineering).



Source: WoS, own calculations and illustrations

Table 35: Proposed rules for Engineering.

Feature & condition	Separated class	New documents		Old documents	
		Classified as new	Classified as old	Classified as new	Classified as old
Fields of Journal = 1	New	45	(72)	10	(118)
Age of Journal < 11 AND Fields of References < 14 AND Countries < 3	Old	(92)	25	(71)	57
Authors > 4	Old	(81)	36	(64)	64
Size of Journal (108;228) AND JIF > 1	New	51	(66)	42	(86)
Age of References > 10	New	80	(37)	73	(55)

Source: WoS, own calculations and illustrations

The first of the conjunctive rules suggests that new documents are published in journals with a broader focus (Table 35). However, a comparison with the descriptive analysis (not shown) suggested that rather the documents published in both extremes of the feature values are new documents. Thus, the rule is extended to journal fields having either the minimum or maximum value of journal fields in the respective year in Engineering.

The next rule consists of three parts which are connected by an “and”-conjunction but will be inspected separately for the purpose of the rule derivation. The first part states that the publications in younger/newer journals are in tendency old documents.

This trend was also observed in the genuine dataset (see above) and the regression model (Chapter 6). However, according to the descriptive analysis, this rule will be changed so that the age of the journal should be 8 or larger. The second part (age of references) cannot be confirmed as the descriptive analysis (and the regression) showed that the new documents are more or less equally distributed among all feature values. Thus, this rule was omitted. The third and last part of the rule however can be made even more restrictive based on the descriptive analysis as 82% of the new documents have authors from only one single country (data not shown). The next rule which concerns the authors can be confirmed as well by the descriptive analysis as the majority of new documents have a number of authors lower than the average.

The rule about the size of the journal as well as the one about the age of the references cannot be confirmed with the genuine dataset and they are thus dropped for further analysis. The remaining rule, stating that the JIF is greater than 1 for new documents applies to 90% of the new documents and is thus tested;

- R1: Age of the Journal ≥ 8
- R2: Country = 1
- R3: Fields of the journal = min or max
- R4: JIF > 1
- R5: Authors < average

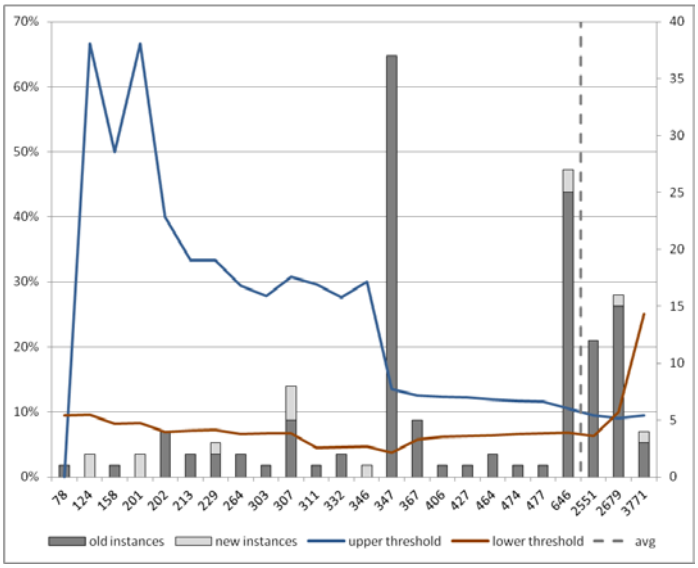
All of the above rules except for R4 proved to be applicable on the Training Set and were thus used in the later testing on the Test Set.

Molecular Biology & Genetics

For Molecular Biology, a huge gap can be found in terms of journal sizes (Figure 47). While 105 of 137 documents have a journal size of 646 items and below, there are some documents that have values of several thousand documents. This makes deduction with the average rather difficult. Nonetheless, the specific journals for this discipline can be closer inspected to make deductions:

- Science (size 2551) published only old documents
- Nature (size 2679) contains 16 documents of which one is new
- The Proceedings of the National Academy of Science of the United States of America (size 3771) covers 1 new and 3 old documents of the dataset

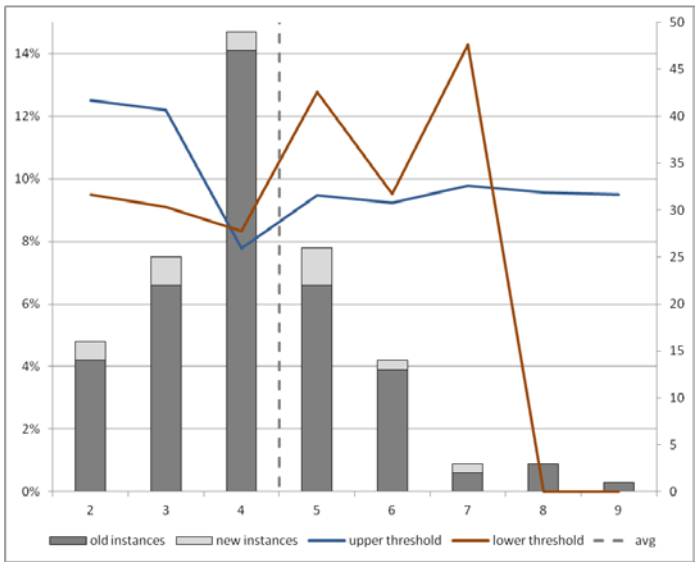
Figure 47: Journal size (Molecular Biology & Genetics).



Source: WoS, own calculations and illustrations

These journals and the respective documents increase the overall average. Also, there is a huge gap between journal sizes 347 and 367: The 37 documents that are excluded based on an upper threshold higher than (or a lower threshold smaller than) 347 are all of established topics and all published in "Nature Genetics". Overall, looking at the journal titles a high focus on Genetics can be found for the emerging topics. In particular, the following journals were used: Genome Research (size 201), PLOS Genetics (229), Genes and Development (307) and Genome Biology (346). Furthermore the above list shows that the major journals with a broad focus publish new documents only in rare cases.

Figure 48: Age of references (Molecular Biology & Genetics).

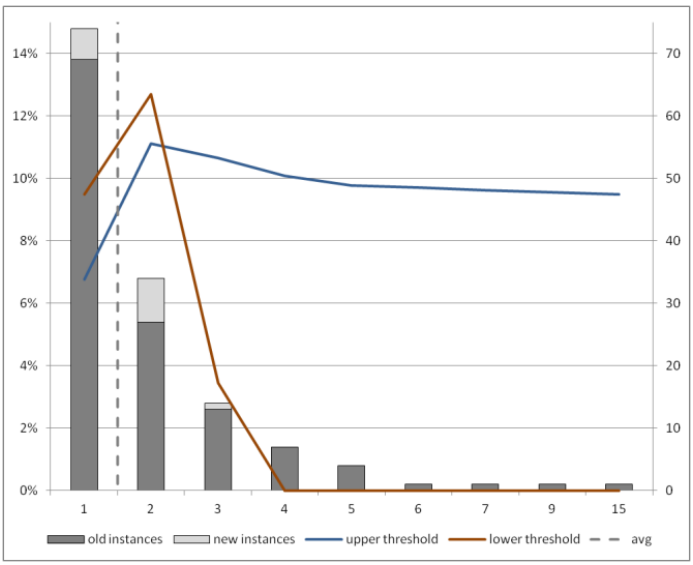


Source: WoS, own calculations and illustrations

In Molecular Biology, the lower threshold shows an increasing trend for new documents with increasing average age of references (Figure 48). Yet, this is a rather slight trend which also is disrupted by the publications with the highest reference age which are all old. Thus, the derivation of a general rule

is difficult. It will be analyzed later if this feature can be used as a stand-alone solution in the proposed approach. For instance, the exclusion of those 21 documents that have a value of 6 and higher seems promising as a supporting rule.

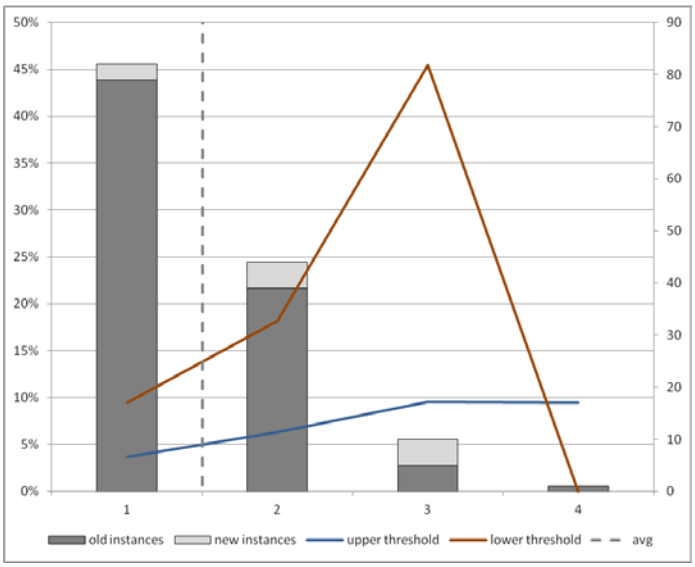
Figure 49: Countries (Molecular Biology & Genetics).



Source: WoS, own calculations and illustrations

More distinctively, Figure 49 shows the trend that new documents are published with only low numbers of author countries. Both threshold show a peak at a feature value of 2, as the majority of new documents in this discipline are published with two countries involved whereas the majority of old documents are published with only one country. Nevertheless, a rule could exclude all documents with a country number of 3 or more because there is only one new document that has 3 countries but no other with more countries.

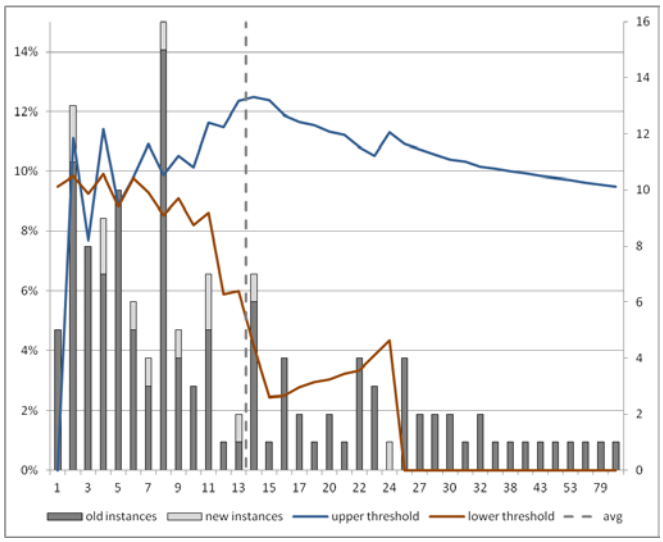
Figure 50: Fields of the journal (Molecular Biology & Genetics).



Source: WoS, own calculations and illustrations

The number of fields in the publishing journals cannot be interpreted so clearly but the lower threshold suggests that – with the exception of the most extreme feature value of 4, which is only observable for one document – the chance for a new document increases with an increasing number of fields in the journal (Figure 50). The limitation to documents which were published in journals with a minimum of two fields excludes 79 old documents while only 3 out of 13 new documents are lost. For a threshold of 3, the ratio is even 5:6 new to old documents, but also the total number of new documents is reduced to a minimum. Thus, a rule that demands more than two fields in the journal for the result set will be tested later. A more general rule for later purposes seems advisable, which can be accomplished by substituting the absolute value with the discipline's average.

Figure 51: Number of authors (Molecular Biology & Genetics).



Source: WoS, own calculations and illustrations

The number of authors in Molecular Biology & Genetics shows a falling trend of new documents for higher numbers of authors (Figure 51). Indeed, such a trend is visible in all topics but Plant & Animal Science. The upper threshold drops for all other disciplines for a higher number of authors. The actual values differ, as do the distributions. In the rules section, it will be confirmed whether this observation can be used for separating the new documents from the old ones. It should be noted that all of the above findings can neither be confirmed nor rejected with the regression model in Chapter 6; for the aggregated discipline, no significant variables could be found.

The rules found by the Conjunctive Rules concerning the size of the journal, the fields of the journal and the number of countries can be confirmed by the descriptive analysis (Table 36), as can the composites of the next to last rule. The age of the journals as well as the JIF cannot be confirmed and are thus omitted from further analysis.

Table 36: Proposed rules for Molecular Biology & Genetics.

Feature & condition	Separated class	New documents		Old documents	
		Classified as new	Classified as old	Classified as new	Classified as old
Size of Journal <347	New	38	(30)	18	(51)
Fields of Journal < 2	Old	(55)	13	(39)	30
JIF > 14	Old	(44)	24	(35)	34
Age of Journal < 7	New	30	(38)	13	(56)
Fields of References < 9	Old	(50)	18	(43)	26
Authors > 14 AND Age of References < 5	Old	(54)	14	(40)	29
Countries > 3	Old	(47)	21	(35)	34

Source: WoS, own calculations and illustrations

The rules tested then were the following:[94]

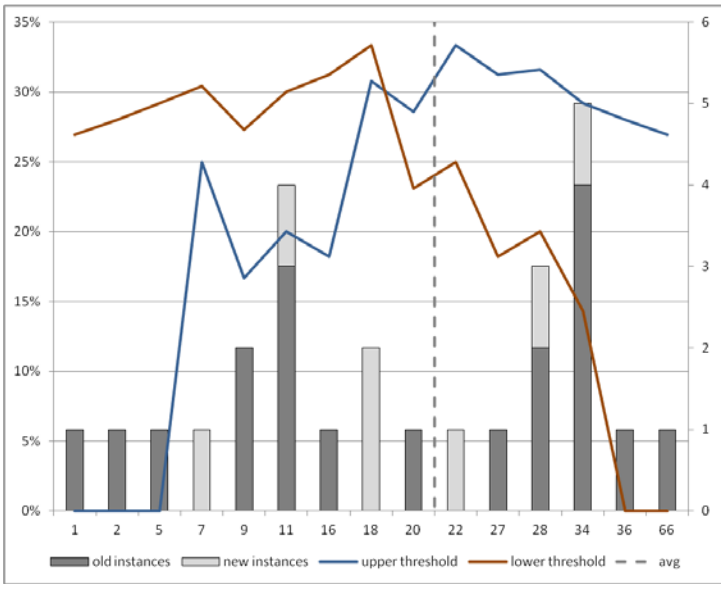
- R1: Fields of the journal >= 2
- R2: Authors <= 14
- R3: Age of References >= 5
- R4: Country <= 2
- R5: Size of the journal < 347

The combination of all rules except R4 performed best on the Training Set and led to a result set that contained 6 new and 5 old documents and a $F_{0.5}$-Measure of 50%.

94 The rules that were contradictory to the descriptive analysis were tested nonetheless, but could not lead to an improvement.

Pharmacology & Toxicology

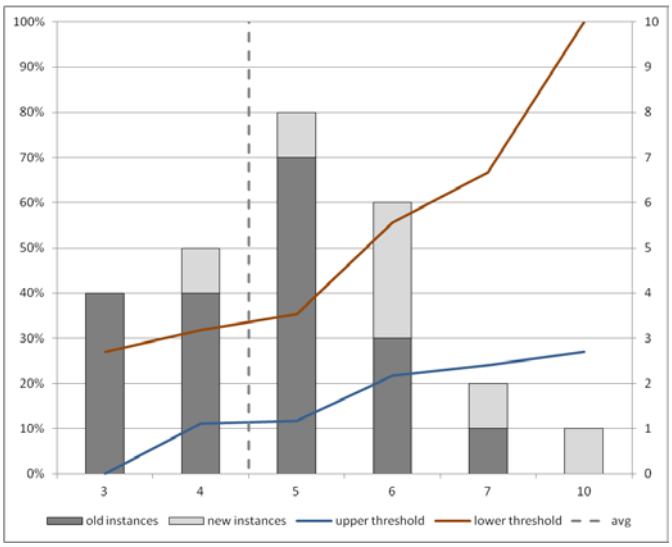
Figure 52: Age of the journal (Pharmacology & Toxicology).



Source: WoS, own calculations and illustrations

The thresholds in Pharmacology and Toxicology show clearly that new documents are neither published in the oldest nor in the newest journals (Figure 53). The problem is how to define the threshold value for the rule as the absolute values might change in different time periods. To be on the safe side, the minimum and the maximum values (7 and 34 years respectively) in the new document set were taken.

Figure 53: Reference age (Pharmacology & Toxicology).



Source: WoS, own calculations and illustrations

For the age of the references, the share of new documents in the subset is increasing for an increasing lower threshold and decreasing for a decreasing upper threshold (Figure 53). All new documents except for one have a reference age value higher than the average. The lower threshold value thus in-

creases from 27% in the initial dataset to 56% for values higher than the average to 100% for the highest value of average reference age in this discipline. Thus, publications in new topics indeed rely more heavily on older publications than those in established topics.

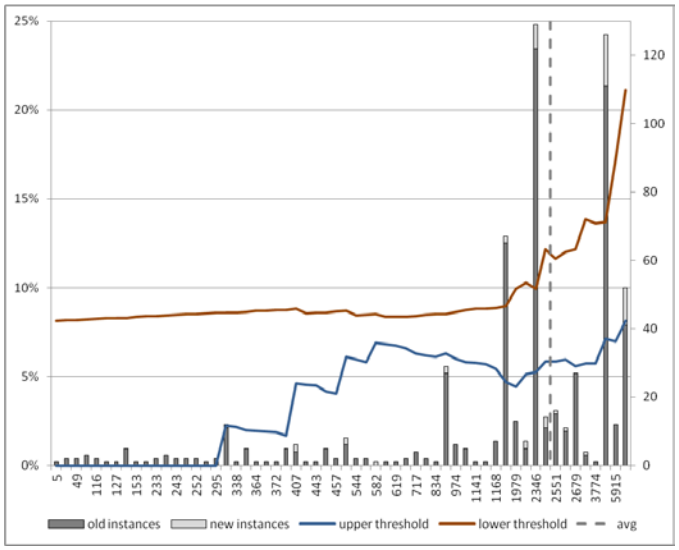
In Pharmacology & Toxicology, the Conjunctive Rule was without result. Thus, the following rules were derived from the descriptive analysis:

- R1: Age of the journal between 7 and 34
- R2: Age of References >= 6
- R3: Authors <= 8 (not shown)
- R4: Country <= 2 (not shown)

The combination of the first two rules led to the highest $F_{0.5}$-Measure value (71%): new documents are by trend published in middle-aged Journals (R1) while their references are much older than the average (R2).

Physics

Figure 54: Journal size (Physics).



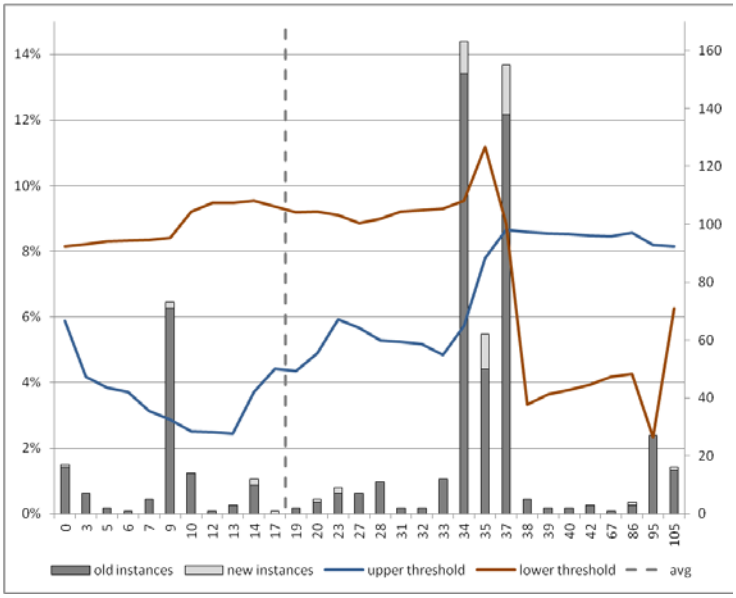
Source: WoS, own calculations and illustrations

The smaller journals in Physics do not show a higher share of emerging topics (Figure 54). On the contrary, all documents with journal sizes below a value of 295 do not belong to a new topic (reflected in an upper threshold value of 0) and approximately three quarters of the new documents were published in a journal with an item size of 1,000 or more. This observation reflects the findings in the regression model for the aggregated discipline (Chapter 6). The journals with journal size 3,816 and 5,916 which cover 26 of the 51 new documents are Physical Review B and Physical Review Letters. The former is specialized despite its size and "devoted to condensed matter and materials physics"[95]

95 http://prb.aps.org/, last accessed on 2013/04/23.

while the latter "provides its diverse readership with weekly coverage of major advances in physics and cross disciplinary developments"[96].

Figure 55: Age of the journal (Physics).



Source: WoS, own calculations and illustrations

Figure 55 is slightly misleading as the thresholds are prone to small absolute values beside the vast majority of articles that are published in journals of the age 34 to 37. These journals also cover the majority of new documents (40 out of 51). Other groups of documents with the same journal age cover two new documents at most. The journals with the ages 34 to 37 are:

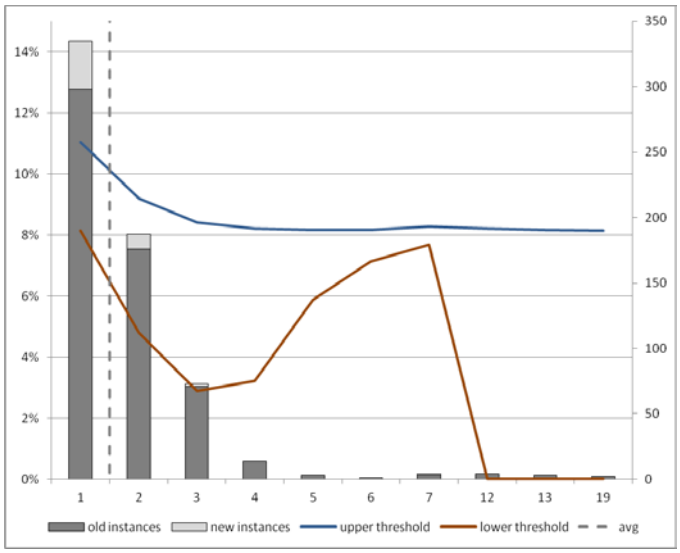
- JETP (age 34, 1 old document)
- Journal of Low Temperature Physics (age 34, 2 old documents)
- Journal of Physics B-Atomic Molecular and Optical Physics (age 34, 2 old documents)
- Journal of Physics D-Applied Physics (age 34, 1 old document)
- Nuclear Physics (age 34, 1 new and 11 old documents)
- Nuovo Cimento Della Societa Italiena di Fisica B-General Physics Relativity, Astronomy and Mathematical Physics and Methods (age 34, 1 old document)
- Physics Review A (age 34, 3 new and 1 old documents)
- Physical Review D (age 34, 7 new and 122 old documents)
- Progress of Theoretical Physics (age 34, 2 old documents)
- Annals of Physics (age 35, 2 old documents)
- Physica Scripta (age 35, 1 new document)
- Physica Status Solidi A-Applications and Materials Science (age 35, 1 old document)
- Physical Review B (age 35, 11 new and 41 documents)
- Physical Review C (age 35, 6 old documents)

96 http://publish.aps.org/about, last accessed on 2013/04/23.

- Physical Review Letters (age 37, 15 new and 111 old documents)
- Physics Letters B (age 37, 2 new and 27 old documents)

This list makes it obvious that one could also restrict the search for new documents in Physics to a set of major journals; the Physical Reviews series A to D and Letters would cover 36, i.e. approx. 71%, of the new documents.

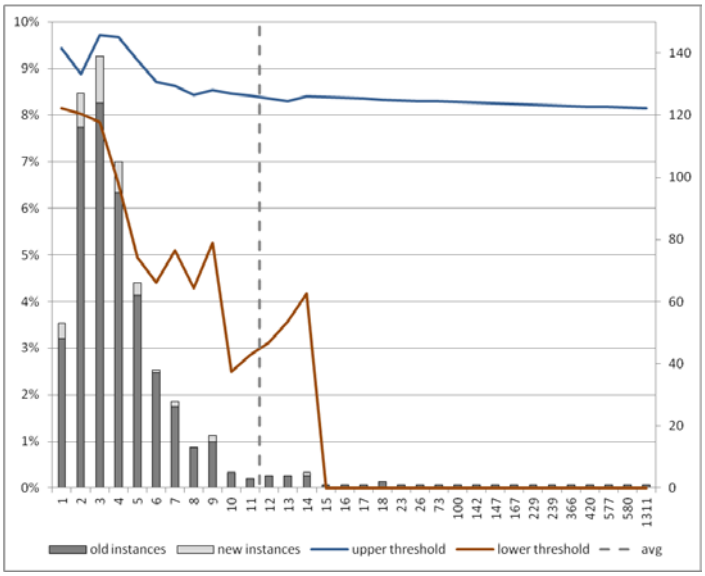
Figure 56: Countries (Physics).



Source: WoS, own calculations and illustrations

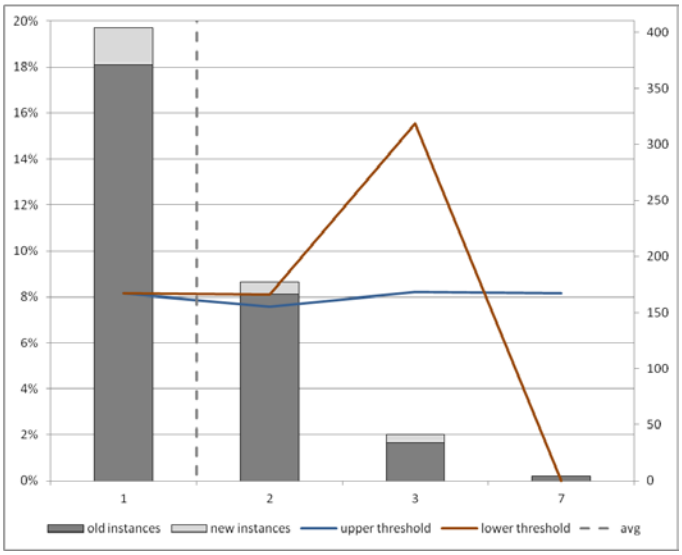
Despite the overall low numbers in Physics, a trend for the countries as expected by the hypothesis is visible in Figure 56. The only exception is one new document with 7 countries that leads to an increase in the lower threshold value. Otherwise, all new documents have a number of countries of 3 or less. In comparison with the other categories, a feature value of 3 is very high, but the overall distribution in countries has the highest variance in Physics. Thus, in relation, 3 countries are comparable to the 1 or 2 country rules in the other disciplines. Physics is also the discipline with the highest number of documents. Therefore, the 2 new documents with 3 involved countries are more or less negligible. The rule that the new documents have two or less countries will be tested later. Furthermore, a similar behaviour can be found for the number of authors (see Figure 57): The higher values are excluded for the new documents.

Figure 57: Number of authors (Physics).



Source: WoS, own calculations and illustrations

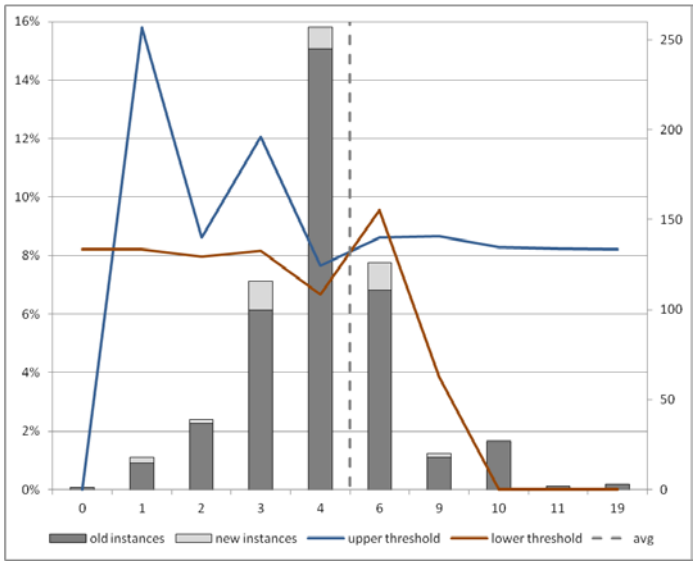
Figure 58: Fields of the journal (Physics).



Source: WoS, own calculations and illustrations

For Physics, Figure 58 shows that the majority of new documents is published in journals with less than average fields. Thus, even though the relative chance of finding a new document (and thus the lower threshold) increases with an increasing number of fields, the majority of them can be found in the more specialized journals.

Figure 59: JIF (Physics).



Source: WoS, own calculations and illustrations

Figure 59 shows that no new documents were published in journals with a JIF higher than 9. The majority of them have been published in Journals with a JIF between 3 and 6. The upper threshold shows that the relative share of new documents is higher in those journals with a low JIF.

Table 37: Proposed rules for Physics.

Feature & condition	**Separated class**	**New documents**		**Old documents**	
		Classified as new	**Classified as old**	**Classified as new**	**Classified as old**
Size of Journal < 393	Old	(206)	86	(190)	127
Age of Journal < 14 AND Age of References > 2	Old	(290)	2	(267)	50
Fields of Journal > 2 AND Authors < 4 AND Fields of References > 6	New	182	(110)	150	(167)
JIF > 3 AND Countries (1;6)	Old	(190)	102	(167)	150

Source: WoS, own calculations and illustrations

Since the reversal of an and-conjunction results in an or-conjunction, the rules presented here can be split into single rules. The rules for the size of the journal, the age of the journal, the fields of the journal, the JIF, the authors and the countries (Table 37) can be confirmed with the descriptive analysis but also adapted to have a more general scope or more precise results. For instance, the journal fields

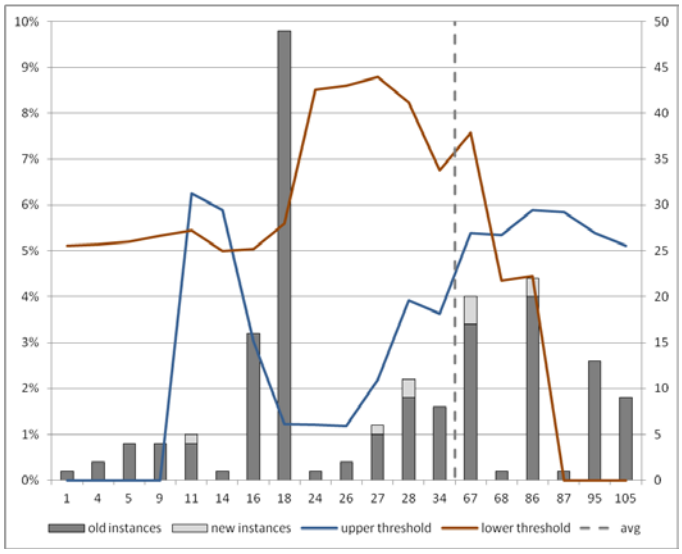
and the number of authors can be set in relation to the average in the discipline. The size of the journal was set in relation to the value 300 to better suit the results gained with the threshold. The rule covering the age of the references was with regard to the number of documents covered less useful. The same was true for the fields of the references. Therefore, the list of tested rules included the following:

- R1: Age of the journal ≥ 14
- R2: Size of the Journal > 300
- R3: Country ≤ 2
- R4: Fields of the journal < average
- R5: JIF between 3 and 6
- R6: Authors < average

In the evaluation of the single rules and their combination, adding R3 and R4 improves the results only slightly in comparison with applying only R1 and R2. However, given the initial size of the dataset, more restrictive rules seem to be appropriate and thus the favour is also given to this. Thus, in the end, rules R1 to R4 are used and lead to a $F_{0.5}$-Measure of 20%.

Plant & Animal Science

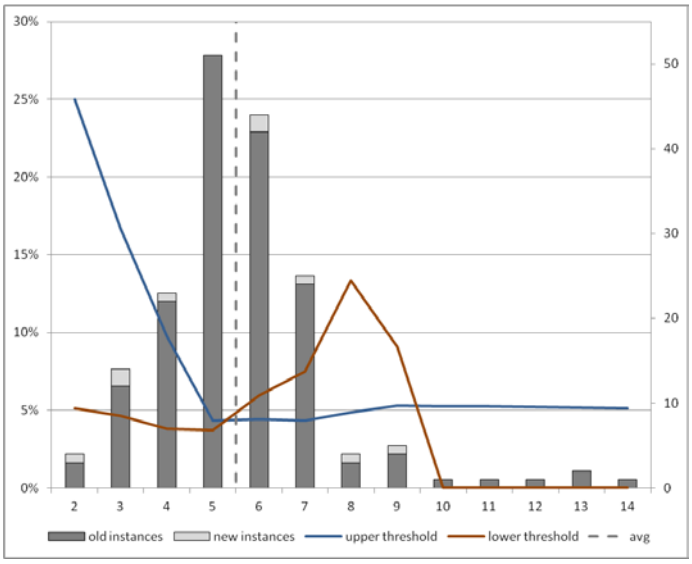
Figure 60: Age of the journal (Plant & Animal Science).



Source: WoS, own calculations and illustrations

Except for one document, all new documents in Plant & Animal Science have a journal age of 27 years or older (Figure 60). Also, the 3 highest values of age (87, 95 and 105) only apply to old documents. According to the graph, most new documents are published in journals have an age close to the average.

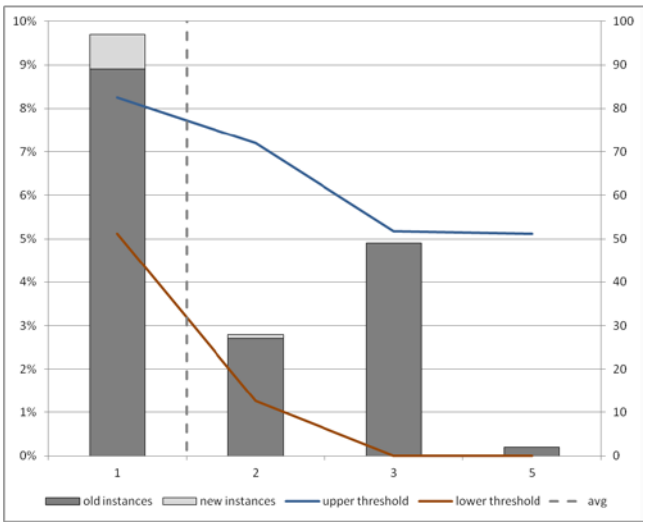
Figure 61: Reference age (Plant & Animal Science).



Source: WoS, own calculations and illustrations

The lower threshold in Plant & Animal Science for the reference age shows that no new documents can be found in the document set with the oldest reference age (Figure 61).

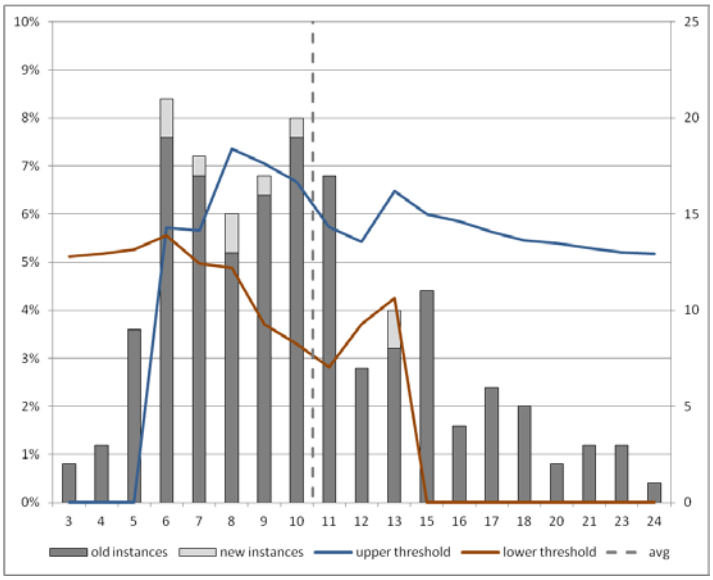
Figure 62: Fields of the journal (Plant & Animal Science).



Source: WoS, own calculations and illustrations

Journals with fewer fields are chosen for the publication of new findings or topics in Plant & Animal Science (Figure 62). Out of the 9 new documents, 8 are published in a journal with one field and only one is published in a journal with two fields. Thus, excluding all journals with more than one field of the journal or a value higher than the average, which corresponds to an upper threshold of 1, would exclude 78 of the old documents but only one of the new documents. Similarly for the number of fields in the references, new documents are published rather with fewer fields. All new documents can be found in journals with 8 to 15 fields. A respective rule will be tested later. The number of fields in the references also varies around the average values whereas the most extreme cases are only old documents (Figure 63).

Figure 63: Fields of references (Plant & Animal Science).



Source: WoS, own calculations and illustrations

Table 38: Proposed rules for Plant & Animal Science.

Feature & condition	Separated class	New documents		Old documents	
		Classified as new	Classified as old	Classified as new	Classified as old
Fields of Journals >2	Old	(34)	53	(1)	88
Authors > 18	New	34	(53)	1	(88)
Fields of References >15	Old	(87)	0	(55)	34
Size of Journal (144;390)	Old	(87)	0	(53)	36
Age of Journal <27	Old	(87)	0	(56)	33

Source: WoS, own calculations and illustrations

The rules for the authors, the size of the journal and the fields of the references were adapted according to the descriptive analysis (Table 38). Rules considering the number of countries and the age of the references were added to account for the observations made in the descriptive analysis. Thus, all in all 7 rules were tested:

- R1: Size of the Journal ≥ 390
- R2: Age of the Journal ≥ 27
- R3: Country = 1
- R4: Fields of the Journal = 1
- R5: Fields of References ≤ average
- R6: Authors ≤ 7
- R7: Age of the references < 8

The appliance on the Training Set confirmed that the new documents are by trend published in highly specialized journals (R4) that are older than for the majority of publications (R2). However, the references are relatively young (R7). The assumption that communication in new topics might be impeded is confirmed, as in most cases all authors stem from one country (R3) and less authors are involved in total (R6).

9.4.3 Derived Features of Emerging Topics

Table 39 shows the rules that were found with the foregoing analysis. These rules were in most cases adapted to use relative instead of fixed values. For instance, the rule for Engineering considering the fields of the journal was that the value equals either 1 or 5, which corresponds to the minimum and maximum values in that discipline.

Table 39: Derived rules for the disciplines.

Discipline	Rules (and conjunction)
Computer Science	Fields of the journal < Average Fields of the references ≥ Average Age of the references between 4 and 8 Journal age ≤ 27
Engineering	Age of the Journal ≥ 8 Country = 1 Fields of the journal = minimum or maximum Authors < average
Molecular Biology & Genetics	Fields of the journal ≥ average Size of the journal < 347 Authors ≤ (rounded) average Fields of the references > 8
Pharmacology & Toxicology	Age of the journal between 1/3 of average and 1/2 of maximum Age of References > average
Physics	Size of the Journal > 300 Age of the Journal ≥ 14 Country ≤ 2 Fields of the Journal ≤ average
Plant & Animal Science	Country = 1 Fields of the Journal = 1 Age of the references < 10 Authors ≤ (rounded) average Journal age ≥ 27

Source: WoS, own calculations and illustrations

9.5 Evaluation

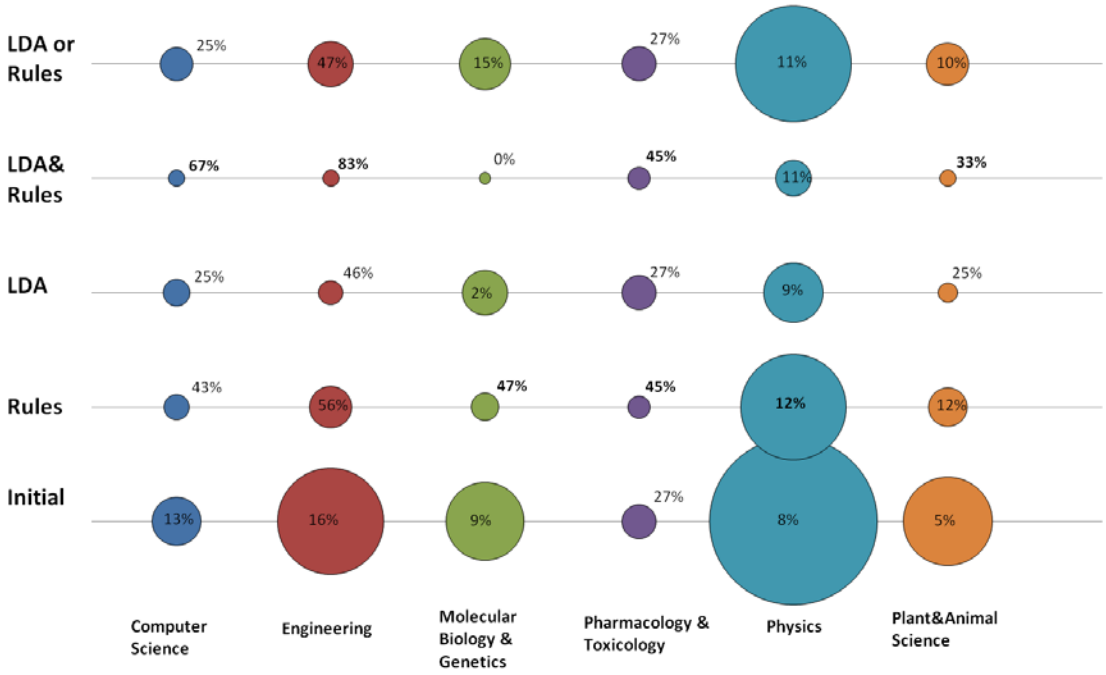
9.5.1 Final Testing on Training Set

LDA, rules and the combination thereof were applied to the Training Set to show the final results that could be achieved with the data on which the parameters were optimized (see Figure 64). The size of the circles represents the size of the respective document set. The figures in percent denote the Precision of the respective document set. For instance, in Computer Science the initial Precision value is 13%. The size diminishes with the application of the approaches while the Precision increases.

In most cases, the combination of LDA and rules performed best as this reduced the result set to a minimum size while showing the highest Precision. In particular, this was true for Computer Science, Engineering, Pharmacology & Toxicology, and Plant & Animal Science. In the latter, the result set consists of only 6 documents, of which 2 were new. The other 163 old and 7 new documents in the initial dataset were sorted out.

In Pharmacology & Toxicology, the rules as a sole solution and in combination with LDA performed equally well. In this case, the result set of LDA covered the result set of the rules but also more old documents. Thus, the only effect of their combination was the restriction to the result set of the rules. In Physics, the rules performed best but also covered still a relatively high number of documents in total (249 out of 626). The results with the combination of LDA & rules are slightly worse but leave only 28 documents for the user to inspect. Only in Molecular Biology & Genetics, the combination of LDA & rules fails. Different publication behaviour or characteristics might hinder transferring the approach, which was tuned on Computer Science, without changes to this or other disciplines.

Figure 64: Size (in documents) and Precision of sets before and after application of the approach on the Training Set.

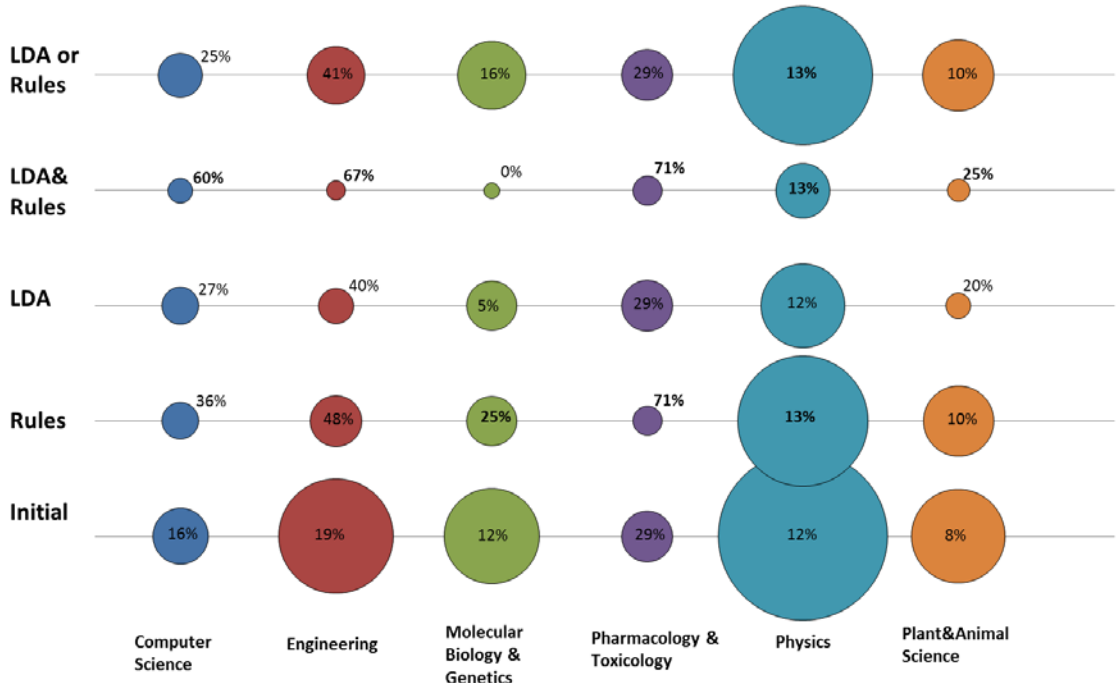


Source: WoS, own calculations and illustrations

Notes: The size of the circles represents the size in documents; the Precision is specified by the numbers in percent. The numbers in bold are the best results for Precision for each discipline.

Similar outcomes can be observed when the results are presented on the topic level (Figure 65). The decline in Precision for Computer Science, Engineering, and Plant & Animal Science suggests that the identified new documents belonged in the majority to the same topic(s). In contrast, in Pharmacology & Toxicology the Precision increases when switching to the measurement of topics. Thus, in this case most documents found by the approach represent different topics. Still, also on this aggregation level the combination of LDA & rules led to the best outcomes with regard to result set size and Precision. Again, only for Molecular Biology & Genetics, the bad transferability of the LDA approach worsened the overall performance.

Figure 65: Size (in topics) and Precision of sets before and after application of approach on the Training Set

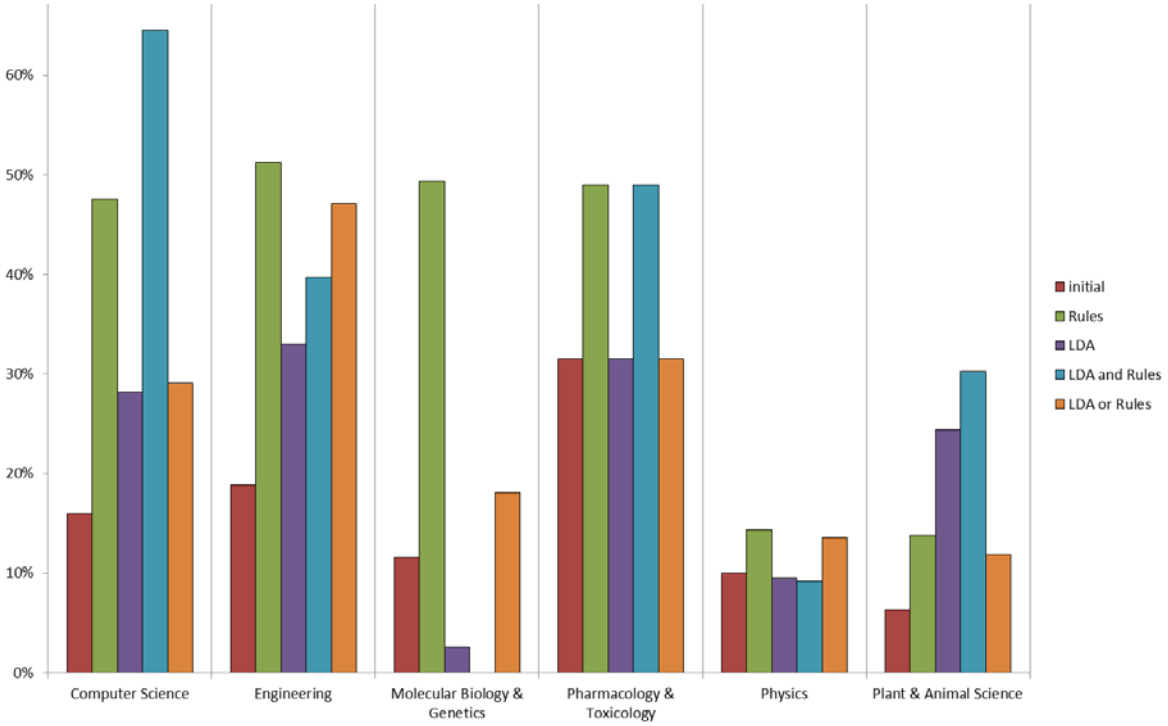


Source: WoS, own calculations and illustrations

Notes: The size of the circles represents the size in topics; the Precision is specified by the numbers in percent. The numbers in bold are the best results for Precision for each discipline.

The results achieved with the $F_{0.5}$-Measure are not as clear as those for the Precision. The losses in coverage caused by the massive reduction in size as shown in Figure 64 reduce the Recall and thus the $F_{0.5}$-Measure. Only in Computer Science, Pharmacology & Toxicology, and Plant & Animal Science the combination of both approaches outperforms the other variants. In Engineering, the combination of LDA & Rules has a higher $F_{0.5}$-Measure value than the initial set, but since the rules work better than the LDA in this case, the sole application of them or an "or"-combination of both approaches leads to better results.

Figure 66: $F_{0.5}$-Measure for the Training Set



Source: WoS, own calculations and illustrations

To show a more consolidated approach for all disciplines, the "and"-combination of LDA and the rules was tested on the Test Set in the following.

9.5.2 Evaluation with Test Set

In the following, the overall approach is tested on the Test Set. Overall approach in this context means the approach as described in the previous section as a combination of LDA and rules.

To test the existing rules on the new disciplines, a mapping had to be done to justify the application. Table 40 shows the mapping that was performed. Computer Science and Pharmacology & Toxicology performed similar in the previous analysis, so both were mapped on Chemistry to test which rule set would be better suited. Computer Science was also used as a basis for the approach for Mathematics, as these disciplines have similar backgrounds. Engineering was used with the same reasoning for both Space Science as well as Materials Science.

Table 40: Transfer of rules to Test Set.

Rules of …	**…were applied on…**
Pharmacology & Toxicology/ Computer Science	Chemistry
Computer Science	Mathematics
Engineering	Space Science
Engineering	Materials Science

Source: WoS, own calculations and illustrations

To better interpret the following results, Table 41 list the number and share of topics in the Test Set (cf. with number of documents in Table 27). Like for the Training Set, the share of new documents/topics in the initial dataset corresponds to the initial Precision. In most cases (except for Materials Science), the initial share of new documents is lower than in the disciplines of the Training Set.

Table 41: Size of Test Set in number of topics (research fronts).

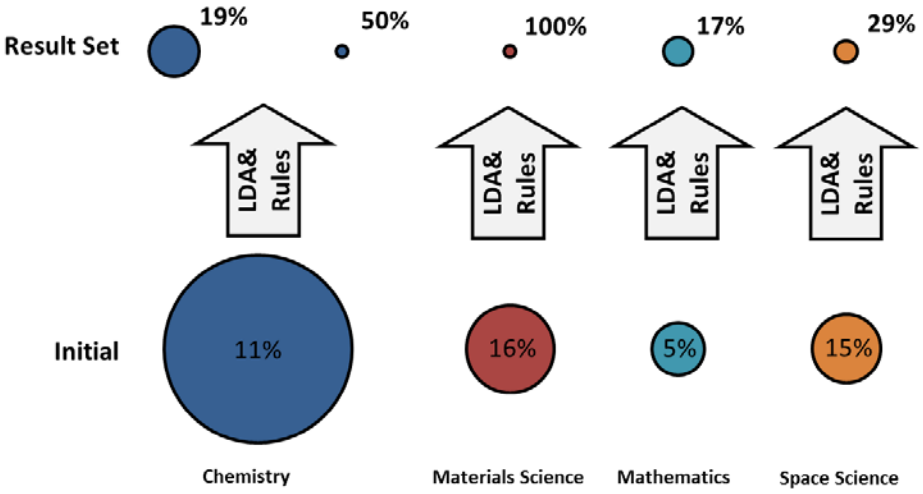
	New topics	Old topics	Total	Share new topics
Chemistry	39	198	237	16%
Materials Science	16	66	82	20%
Mathematics	1	19	20	5%
Space Science	8	24	32	25%

Source: WoS, own calculations and illustrations

The respective rules and the LDA approach were applied on the different disciplines separately. Please note that while LDA used the same parameters for all disciplines, the rules were based on the findings for the disciplines in the Training Set.

Figure 67 shows the results for the single disciplines before and after the approach was utilized. Since two different kinds of rules were applied on Chemistry, two result sets are depicted here: The one on the left was achieved with the rules of Pharmacology & Toxicology, the one on the right with those of Computer Science. Further tests show that the results of rules & LDA and rules are equal in Space Science, i.e. LDA has a result set that covers that of the rules completely. In general, the approach performs differently depending on the disciplines. As already seen in the case of Molecular Biology & Genetics above, the transfer of the approach from one discipline to another can influence the performance. Even though the transfer was selected based on common discipline characteristics and "behaviour", rules that apply in one discipline do not necessarily hold in another. However, the result sets are surprisingly small as the total numbers listed in Table 42 suggest.

Figure 67: Size (in documents) and Precision of sets before and after application of approach.



Source: WoS, own calculations and illustrations

Notes: The size of the circles represents the size in documents; the Precision is specified by the numbers in percent. Chemistry has – depending on the applied rules – different result sets. The one on the left hand side was achieved with the rules of Pharmacology & Toxicology, the one on the right hand side with those of Computer Science.

Table 42: Initial and result set size.

Discipline	Initial number of documents	Number of documents to inspect
Chemistry – Ph	503	36
Chemistry – CS	503	2
Materials Science	110	2
Mathematics	40	12
Space Science	67	7

Source: WoS, own calculations and illustrations

Notes: Chemistry has – depending on the applied rules – different result sets. "Ph" denotes the one achieved with the rules of Pharmacology & Toxicology, "CS" the one with those of Computer Science.

When interpreting the results as a whole, the approach improves the document set in any case, i.e. the document set size is considerably reduced while the Precision is increased. In Materials Science, the result set merely consists of 2 documents that are both new. In Chemistry, the rules of Computer Science work better with regard to both Precision and size reduction. Again, the Recall is influenced tremendously by this reduction in size and accounts for only 2% in the case of Computer Science. In contrast to that, the result set for the Pharmacology & Toxicology rules covers 7 instead of 1 new document and thus at least has a Recall value of 13%.

Table 43: Normalized comparison of results in the single disciplines.

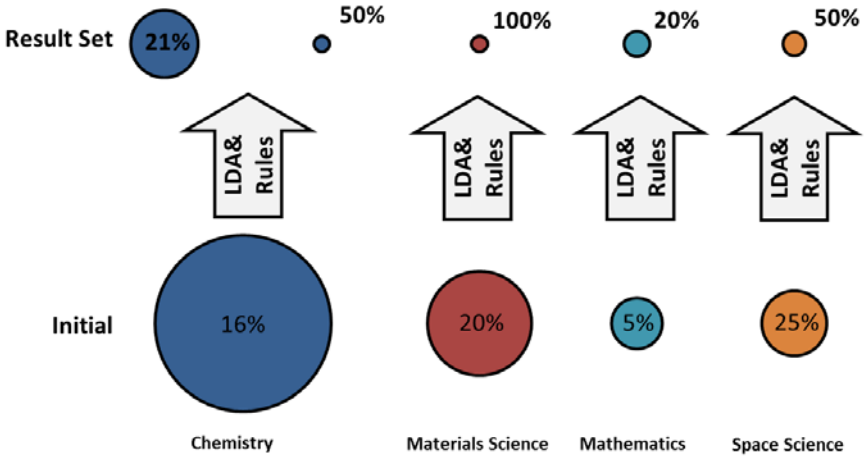
Discipline	Total	New	Old
Chemistry – Ph	72	14	58
Chemistry – CS	4	2	2
Materials Science	18	18	0
Mathematics	300	50	250
Space Science	104	30	75

Source: WoS, own calculations and illustrations

Notes: Chemistry has – depending on the applied rules – different result sets. "Ph" denotes the one achieved with the rules of Pharmacology & Toxicology, "CS" the one with those of Computer Science.

It is difficult to compare to the results in the disciplines with each other, in particular because they have different document set sizes, and furthermore to imagine the outcomes in a real application. Therefore, the results are also represented in a normalized view (Table 43): If the initial set of each discipline covered 1,000 documents, how many documents were returned and how many of them were new/old? The normalized representation better illustrates the performance of the approach even though a real evaluation of that size cannot be performed. Mathematics clearly has the biggest result set with many old documents., but also the highest Recall of 100%. The other result sets could be inspected manually with justifiable effort. In particular in Chemistry and Materials Science, the rate of new documents in the result set is promising.

Figure 68: Size (in topics) and Precision of sets before and after application of approach.



Source: WoS, own calculations and illustrations

Notes: The size of the circles represents the size in documents; the Precision is specified by the numbers in percent. Chemistry has – depending on the applied rules – different result sets. The one on the left hand side was achieved with the rules of Pharmacology & Toxicology, the one on the right hand side with those of Computer Science.

Looking at the number of topics, the results improve even more for some disciplines (Figure 68). In particular for Space Science the Precision increases from 25% to 50%. In Chemistry with Pharmacology & Toxicology rules (left hand side) and in Mathematics, the results improve slightly in comparison with the document numbers. In Chemistry with Computer Science rules (right hand side) and in Materials Science this change of view offers no leeway for improvement and thus the results stay the same. Table 44 lists the total number of topics before and after the application of the approach. An inspection of 2 to 5 groups of documents seems fairly manageable for a human user.

Table 44: Size of topic set (research fronts) before and after application of approach.

Discipline	Initial number of topics	Number of topics to inspect
Chemistry – Ph	237	34
Chemistry – CS	237	2
Materials Science	82	2
Mathematics	20	5
Space Science	32	4

Source: WoS, own calculations and illustrations

Notes: Chemistry has – depending on the applied rules – different result sets. “Ph” denotes the one achieved with the rules of Pharmacology & Toxicology, “CS” the one with those of Computer Science.

Next, the set of new documents that were lost during the process was inspected. Since the Recall of the combination of the overall approach highly depends on the single approaches’ performance, the first step was to analyze their individual performance on the new dataset. Therefore, the share of new documents, that were excluded by each approach and their combination was calculated, where 100% accounted for all of the new topics available in the respective discipline. Table 45 shows that in all cases but Space Science, the share of new documents that was lost was higher for the rules than for

LDA. The combination of both approaches can only worsen the results, as in this way all publications that are excluded (falsely) by one of the two approaches are excluded. However, this step is still justified by the high Precision that can be achieved by this. Mathematics, which was previously the discipline on which the worst performance in terms of Precision were measured, shows the highest Recall here, as no false negative documents can be observed.

Table 45: Exclusion rate of new documents of the two approaches and their combination.

Discipline	Rules	LDA	Rules & LDA
Chemistry – Ph	77%	53%	87%
Chemistry – CS	92%	53%	98%
Materials Science	72%	67%	89%
Mathematics	0%	0%	0%
Space Science	60%	80%	80%

Source: WoS, own calculations and illustrations

Notes: Chemistry has – depending on the applied rules – different result sets. "Ph" denotes the one achieved with the rules of Pharmacology & Toxicology, "CS" the one with those of Computer Science.

Since the set of rules has the higher share of false negatives, the specific rules that caused the exclusion of the new documents were analyzed. Table 46 shows the share of documents which were excluded due to a rule. Of course, the specific rules differed for the disciplines. However, the comparison shows which rules could not be transferred to the new disciplines.

Table 46: Percentage of wrongly excluded new documents and the rule which excluded them.

Disciplines	Journal fields	Journal size	Journal age	Authors	Reference fields	Reference age	Country
Chemistry – Ph	0%	0%	73%	0%	0%	80%	0%
Chemistry - CS	22%	0%	69%	0%	57%	47%	0%
Mathematics	0%	0%	0%	0%	0%	0%	0%
Space Science	0%	0%	0%	50%	0%	0%	100%
Materials Science	46%	0%	31%	62%	0%	0%	38%

Source: WoS, own calculations and illustrations

Notes: Chemistry has – depending on the applied rules – different result sets. "Ph" denotes the one achieved with the rules of Pharmacology & Toxicology, "CS" the one with those of Computer Science. Values higher than 50% are marked in bold.

When applying the Pharmacology & Toxicology rules on Chemistry, many (30) of the 41 misses are a result of the journal age. Probably the mapping via the average and maximum value leads to a misclassification. In addition to that, in most cases (33 out of 41), the reference age is too small and thus the respective documents are excluded from the result set. When applying the rules from Computer Science, the journal age is again the major issue. Here as well, the average values for the two disciplines, Computer Science and Chemistry, differ strongly: In Chemistry, the average is around 44, in Computer Science approximately 27. This might indicate that the overall distribution and relevance of the journal age differ and thus the rules cannot be that easily transferred.

In Space Science, the number of countries involved in new topics is clearly higher than in Engineering; the same is partly true for the number of authors. The higher costs of research projects in Space Science might demand funding across country boarders or work in bigger research teams. Also, the results in the descriptive analysis showed that the number of countries is higher in Physics (see Figure 56), which might also reflect here.

In Materials Science, most missed new documents are caused by differing values in journal fields and too many authors per paper. According to the rule from Engineering, the number of fields in the journal should only take the maximum and minimum value in the discipline, in the case of Materials Science this would correspond to values of 1 and 6. However, 6 new documents were published in journals with two research fields, which seems like a rather small deviation.

Note that in most cases, problems occurred even though no fixed value was used but the average of the discipline. Thus, changing the absolute values to relative ones did not necessarily improve the transferability of the rules.

9.6 Summary

This chapter presented the calibration, composition and evaluation of the proposed approach. With the help of different Training Sets, parameters for the LDA approach and the similarity calculation were estimated and rules for the outlier detection were derived.

Best results were achieved on the Training Set with the combination of both approaches, LDA and rules. Therefore, their combination was used on the Test Set as well. Since the rules differed for the disciplines, a mapping between the disciplines in the Training Set and those in the Test Set had to be found. In the end, the rules of the disciplines Pharmacology & Toxicology, Computer Science and Engineering were tested on the four new disciplines Chemistry, Mathematics, Space Science and Material Science. Only with this mapping, could the evaluation be conducted independently from the training data. The content-based aggregation however could still lead to the transfer of rules between disciplines that have different publication behavior.

Best results in terms of Precision were achieved in Chemistry and Materials Science with values of 50% and 100%. Mathematics performed worst, but further investigations showed that the Recall in this particular result set accounted for 100%. In most cases, documents were lost because the values or thresholds for the new documents seemed to differ slightly. For instance, in Space Science the number of countries was higher than in Engineering in general.

In all disciplines, the size of the set for manual inspection could be reduced to 0.4% to 30% of the initial documents set size. In combination with the high Precision values, the results seem promising for future applications. In the following chapter, the application of the approach on another dataset is shown, which enables a better understanding of the term development.

10 Emerging Topics – How New Terms are Introduced in the Scientific Landscape

10.1 Introduction

A main part of the approach presented in this thesis is based on the assumption that a topical change is pronounced by a different usage of terms. In particular, it is assumed that topics use terms in varying quantity and that new topics introduce a new vocabulary or use existing terms in novel combinations. The identification of emerging topics mainly focused on the detection of new vocabulary and new combinations of existing terms. It has already been noted that – on the level of document sets – "words change both in terms of frequencies of relations with other words" (Leydesdorff 1997, p. 418). In this final chapter of the thesis, the representation of changes in the scientific landscape via term vectors is tested. Therefore, the approach is applied on an original dataset. The resulting emerging topics enable the study of the term usage. The analysis compares the vocabulary used at the point of the emergence of at topic with the distribution in the overall dataset. In this way, new structures in the vocabulary can be identified. Also newly introduced terms can be found. The exemplary findings support the high focus on term usage in the approach presented in this thesis.

The evaluation so far concentrated on replicating other (older) results. This stemmed from the fact that only with the help of such existing Gold Standards, an automatic and structured evaluation was enabled. In addition to the term analysis, the application of the approach on a new random dataset from WoS in this chapter was meant as a final evaluation that also delivered content-wise results. Now, the topics detected by the approach in a real-world application were assessed. Again, this was partly curiosity for content-wise results, but furthermore another (intellectual) form of evaluation.

10.2 The Emerging Topics in Focus

This section explains the adaptation of the so far developed approach on the new dataset. The changes were necessary to enable the analysis on a larger scale. Also, the topics found by the approach are presented. Not all of them are depicted in detail. Yet, the content of these topics is explained in a way that is also understandable for non-professionals. The set of clusters presented here was selected respectively – a fact that is explained and considered when the cluster selection and the results are discussed in Sections 10.2.2 and 10.3.

10.2.1 Identification via Introduced Approach

Changes had to be made to adapt the approach for the new dataset. In particular, since the approach had so far been only parameterized and evaluated on relatively small datasets, changes in regard of the scalability had to be made. While the overall procedure stayed the same, the following parts of the implementation, which will be discussed in more detail below, were adapted:

1. Accessing the SQL Oracle database via Java Database Connectivity (JDBC)
2. Using an enhanced list of stopwords
3. Selecting dominant topics in consideration of ambiguous topics
4. Connecting topics with an adapted threshold parameter t_c

The first change concerned the now necessary interface with JDBC. It facilitated the data access in particular for larger datasets. In this way, datasets could be defined more flexible, while the local storage space could be reduced. By this means, data were requested in real-time and did not have to be extracted from the bibliometric database in advance. Overall, this was also a step to facilitate a later integration in other systems or an adaptation for other purposes.

The second point considered new terms that were found by a thorough check of the abstracts in the new dataset. Frequent terms in these abstracts were sometimes "residues" from the database processing (an example for that is given below) or frequent terms specific for scientific publications. The latter are for instance terms that are used more often by researchers in abstracts or publications without adding to the content.[97] For example, terms like "recently" etc. are not helpful for the delimitation of topics. The approach was more prone to be disturbed by such "scientific stopwords" when applying it to a larger dataset. Thus, the list of stopwords was first extended with a set of terms that could be found frequently in the abstracts at hand and also with those which were included in the list of frequent adverbs in scientific publications with a publication count of two or more.[98] All terms were included in the stopword list once in the adverbial and once in the adjective form. Other terms were covered spuriously in the abstracts. This concerned in particular terms like copyright statements etc. One frequent phrase for instance was the copyright statement by Elsevier ("© 2013 Elsevier Ltd."), which was appended to the original abstract text in the respective documents. Such terms were collected as well manually in the bibliometric datasets and added to the list of stopwords. The fixed threshold t_w, which was used before to exclude frequent terms in general, was replaced by this more precise method.

The third step was necessary, as a phenomenon occurred that also was not apparent in the smaller datasets; the topic model also covered more general topics. These topics were on a higher hierarchical level than others and thus more ambiguous. In turn, they also provided the scientific background for many other topics. Due to their more general definition, these topics prevailed in the dominant topic selection process. Few (ambiguous) topics in one year were connected to the majority if not all of the topics in the following year. This in turn influenced the identification of the emerging topics. Because of that, an additional step in the dominant topic assignment was added which excluded by and by those topics that covered "too many" of the documents at hand. A threshold of 50% was used to identify these nearly omnipresent topics. The overall procedure for the determination of dominant topics was then as follows:

1. For each document, the dominant topic is determined as described in Section 5.2.
2. For each topic, the share in documents for the overall dataset is calculated:
   a. If one topic covers more than 50% of all documents, it is added to a "blacklist" of ambiguous topics. The procedure starts then all over (Step 1) with no assignment of a document to a topic on the "blacklist"
   b. If no topic is added to the blacklist, the procedure finishes.

---

97 For an analysis of trends in frequent terms in scientific (biomedical) literature see http://nsaunders.wordpress.com/2013/07/16/interestingly-the-sentence-adverbs-of-pubmed-central/, last accessed 2014/01/30.

98 https://github.com/neilfws/PubMed/blob/master/adverbs/output/adverbs.freq.csv, last accessed 2014/01/30.

For the theoretically possible case that no dominant topic could be found for a document, a topic of the blacklist was assigned. However, in order for this to happen, the probability for all other topics had to be 0.

One final adjustment concerned the threshold parameter t_c. The connections were more difficult to establish after the explicit exclusion of ambiguous topics. Also, the higher number of overall topics might influence the similarity calculation: The similarity value between clusters with a higher granularity in the topic representation can be expected to be lower than that for general topics. Again, this can be attributed to the shared vocabulary, as the more specific topics might have fewer words in common than those on a higher aggregation level. Thus, the similarity value used for the establishment of connections might be lower and thus less often above the necessary threshold value. With a sample dataset for the years 2000 and 2001, the new parameter for the threshold t_c was determined. For the sample set, the number of clusters without connections, i.e. the apparent emerging topic candidates, was measured. For the previous value of $t_{c,}$ all documents were marked as emerging. Thus, the value was adjusted by and by to yield a more exclusive set of emerging topic candidates. The final value was set to 0.42, with which 2 out of 81 non-empty clusters were marked as emerging topics in the sample dataset. However, the analysis showed that the threshold value was highly dependent on the year on which the approach was applied.

Apart from these changes, the procedure remained the same: First, the LDA procedure was applied to the dataset, then the unconnected clusters were extracted as emerging topic candidates. A dataset covering the journal articles in Artificial Intelligence in the WoS between 2002 and 2010 was used. Only documents that had at least an abstract and a title and one reference were included. For an analysis in the scope of this thesis, the dataset had to be restricted to a maximum of 5,000 documents per year. Each year thus covered exactly 5,000 documents. The year 2003 was used as a starting year only to identify the emerging topics in 2004. The other years were used to derive additional parameters for the term usage. In particular, the distribution of terms in the dataset in different years was compared. For that purpose, an annual ranking was calculated for the overall dataset. Also, the usage was estimated via the percentage of documents covering a specific term in one year. The increase in relative share, i.e. the relative growth, per year was calculated to detect trending terms. This helped to estimate whether a high term usage in a topic was the result of an overall trend or an individual development.

The analysis of term usage was restricted to the ten most common keywords per cluster. The keywords were used in contrast to the terms extracted for the LDA approach to provide a second view on the data: First, the usage of the same terms that were used for clustering would have led to tautological results. The documents clustered together by the approach would necessarily share the same terms. However, the keywords selected by the authors provide another perspective in that they show how the authors would have described and summarized their work with single selected terms. Thus, if similar usage in keywords related to the overall topic of a cluster could be shown, the approach would be corroborated. Second, the keywords could be better related to the overall dataset as this selected set of terms was more often used than other terms in e.g. the abstracts. Third, on a more practical notion, the term vectors as a result of the LDA approach were a mixture of vocabulary and author names, so that a

structured analysis and a comparison with the remainder of the dataset would have yielded only vague results.

The LDA approach was run on the dataset to detect emerging topics in 2004. In this way, 28 topic clusters containing 960 documents were identified as emerging topic candidates. The rules derived in Chapter 9 were applied to reduce the document set to 172 emerging topic candidates. The number of assigned topic clusters was not reduced by that step, i.e. at least one document was selected from each topic cluster as an emerging topic candidate.

The next subsection presents the topics identified by the approach. The content of the topic clusters is explained in more detail, so that the identified terms can be brought into context. These terms are then presented and analyzed in the next section.

10.2.2 Description

The analysis started with a manual assessment of the documents for each cluster and a labelling based on common contents. For 7 clusters, no such common ground could be found so that they were excluded from the further analysis. This might have been an effect of the selection of emerging topic documents, which left the respective document clusters with only – seemingly – unconnected documents, or a failure in the manual assessment. In any case, 21 clusters were left for the remaining analysis. 10 clusters were selected, for which on the one hand a description could be given without delving too deep in Artificial Intelligence foundations. On the other hand, these clusters showed prominent results in the term analysis. The limitations that this pre-selection induced for the overall analysis are discussed in the final section. However, this step was necessary to keep the amount of information manageable for this chapter.

In the following a brief overview of the single clusters is given (Table 47). As in particular the documents identified as emerging topic candidates were of interest, the description (as well as the cluster labelling described above) focused on them. The number of other documents is however given to show the relative size of the clusters.

Table 47: Exemplary set of 10 clusters found in the WoS in the year 2004 and their content.

Title: Support Vector Machines, Text Mining and Machine Learning for gene selection (in particular cancer related)		
ID: 0	**Pub: 101**	**Pub ET:25**
Description: The documents describe various methods from Artificial Intelligence that can be used for gene selection. A particular focus of some of the new documents lay on the identification of cancer related genes.		
Title: Data series analysis for prediction		
ID: 18	**Pub: 29**	**Pub ET: 10**
Description: Using time series of data to predict future developments (also called time series analysis) has a long history in Artificial Intelligence. However, with the introduction of Grid and later Cloud Computing, the processing of data on a larger scale was possible and enabled trend estimation for larger and more complex datasets. In between the emergence of those two hardware-related methods, publication counts for trend analysis might have experienced a surge and evoked topics as the one listed here.		

Title: Adaption for trade-off preferences

ID: 21 **Pub: 31** **Pub ET: 5**

Description: User preferences can be applied in various contexts in Computer Science. However, in most methods, the preferences show little flexibility and interaction. Trade-off preferences concern relationships between preferences that cannot be expressed by standard rules. Cluster no. 21 was concerned with the adaption of existing methods for these purposes.

Title: Monitoring granular changes

ID: 22 **Pub: 24** **Pub ET: 6**

Description: Publications in this cluster were primarily concerned with graphical analysis. Images of temporal data were compared in order to detect changes. The topic also covers the estimation of "risk types", which goes hand in hand with the monitoring of developments.

Title: New methods for optimization in clustering

ID: 42 **Pub: 26** **Pub ET: 6**

Description: This cluster was specifically concerned with the optimization in existing clustering methods. Mathematical or computational methods like numeric or Boolean optimization are applied that provide a common background to the documents associated with this topic.

Title: Constraint representation for pattern recognition

ID: 45 **Pub: 13** **Pub ET: 3**

Description: The main focus of this cluster lay on the recognition of patterns. The similarity in the publications was that all methods relied on the usage of constraints to delimit the patterns from "common noise".

Title: Object delimitation in (moving or static) images

ID: 46 **Pub: 130** **Pub ET: 13**

Description: In the past decade, the automatic identification of specific objects or living beings in visual data gained more and more importance. Thus, a cluster like cluster no. 46 is no rarity. However, its main focus lay on the identification of the boundaries of objects instead of their (fuzzy) detection.

Title: Converting vague descriptions to specific ones

ID: 81 **Pub: 19** **Pub ET: 5**

Description: A further topic with an increasing popularity in Artificial Intelligence in the past decade was natural language processing (NLP). In general, NLP is concerned with the automatic processing of textual or audio input in usual grammatical structure and phrasing. The cluster no. 81 dealt with the conversion of vague descriptions in the form of NLP input or already postulated constraints to more specific rule sets.

Title: Stability of fuzzy systems

ID: 86 **Pub: 47** **Pub ET: 8**

Description: Fuzzy regulations are important to draft relative rules that regard various contexts to form a decision. However, because of the high interaction between the concerned features, the resulting system can behave volatile. In this cluster, the relation between the fuzzy rule set and the system stability were tested.

Title: Collaborative retrieval

ID: 99 **Pub: 19** **Pub ET: 4**

Description: As stated above, new methods to combine hardware power from different sources were introduced in the past decade. This also demanded means to combine the results from different sources. The current cluster was concerned with the requirements when multiple agents for search tasks were used. In particular, different methods were presented to process the search query and combine the various results.

Source: WoS, own calculations and interpretations

Notes: The ID is the cluster ID given by the approach developed in this thesis, "Pub" and "Pub ET" denote the number of overall and new publications in each cluster. The titles were assigned manually.

10.3 Term Development

In this section, the usage of terms in the topics and in the remainder set of scientific publications is compared. In this way it can be determined, which terms were only used because of a high general dispersion at that point of time. The goal is thus to identify those terms that are more or less exclusive in their usage for the topic at hand.

Table 48: Keyword ranking and distribution for Cluster no. 0 ("Support Vector Machines, Text Mining and Machine Learning for gene selection (in particular cancer related)").

Keyword	**Term rank in…**			**Term usage in whole dataset in the year…**								
	ET cand.	**cluster**	**dataset**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Discovery	1	5	128			*					*	
Gene-expression	2	11	435		*	**		**	*			*
Bayesian networks	3	17	76			*						
Cancer	4	3	228	*		*		**		*		
Search	5	22	50									
Selection	6	25	56									
Prediction	7	8	55		*		*					
Classification	8	1	4									
Estimation of distribution algorithms	9	46	737			*			*			**
Molecular classification	10	32	2569	*		*	*			*	*	

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to previous the year), ***: growth, so that usage at least quintuples (with regard to the previous year).

Table 48 shows the specific keywords for Cluster no. 0. The strong focus on biomedical applications becomes apparent by the terms "gene expression" and "cancer". Both terms have a relatively high ranking in Artificial Intelligence in general. However, they were not so widespread until the year 2004, in which an increase in their usage can be noted. Furthermore, the table lists general terms like "discovery", "search", "selection" and "prediction". As these terms are of universal importance for Artificial Intelligence, they bear no implications for this specific topic.

The last two terms in the table seem to be more specific for the topic. These terms are low ranked in the overall dataset. The term "molecular classification" spreads in 2004 for a two years' phase. Overall, an increase in the term usage for the topic specific terms is observable. For instance, "estimation of distribution algorithms' more than doubles it dissemination in 2010. Thus, the topic not necessarily uses only terms that are common at that point in time anyway, but also uses specific terms that induce a later trend in the field.

Table 49: Keyword ranking and distribution for Cluster no. 18 ("Data series analysis for prediction").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Models	1	36	11						*			
Qualitative spatial representation	2	28	3101			*		*				
Queries	3	38	649				*		*			*
Regions	4	39	491		*			*			*	
Networks	5	43	13									
Pathways	6	44	12182			*			*	*		
Algorithm	7	2	3									
Mechanics	8	46	1666	*		*		*	*		*	*
Heuristics	9	50	129			**			*			*
Hypotheses	10	51	928			*			*		*	

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

For Cluster no. 18, a high rank for the keyword "qualitative spatial representation" can be seen, which is seldom used in other contexts (Table 49). This is also the only term in the ranking that is so ambiguous that it can be used in other contexts. Yet, the term distribution never spread across the whole dataset. "Time series prediction" was at this point in time, in the year 2004, not very widely used yet and is therefore not listed in the table. After a first peak in 2002 with 1.4%, the shares of "Time series prediction" in tendency diminished until 2007, wherein a second peak with 2.2% was reached. This observation in combination with the term usage in Cluster no. 18 suggests that the awareness of such specific vocabulary had not spread yet. A closer look at the publications revealed that in that specific case the authors generated their own vocabulary to express the notion of their work.

Table 50: Keyword ranking and distribution for Cluster no. 21 ("Adaption for trade-off preferences").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Model	1	3	2									
Fuzzy	2	66	370				*		*			
Deriving priorities	3	25	2240			*						*
Similarity	4	94	60			*		*				
Regression	5	93	44									
Network	6	72	53						*			

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

Cluster no. 21 turns out to be a very specific topic, which in particular is characterized by a new combination of existing terms (Table 50). Also, only few terms are used, so that the ranking covers only 6 terms for the keywords in the emerging topic candidates. One term in particular, “deriving priorities”, seems to be very particular for this problem. Yet, it was not established in the overall vocabulary set. Thus, it can be concluded that this topic-specific term was introduced in the emerging phase, but could not spread in the scientific field precisely because of its specificity.

Table 51: Keyword ranking and distribution for Cluster no. 22 (“Monitoring granular changes”).

Keyword	**Term rank in…**			**Term usage in whole dataset in the year…**								
	ET cand.	**cluster**	**dataset**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Optimization	1	69	12						*			
Granularities	2	72	10776		*							
Ant colonies	3	67	11341			*		*			*	
Databases	4	55	43									*
Protein	5	51	910	**		*				*		
Cancer	6	45	228	*		*		**		*		
HPV	7	42	12944			*						
E6	8	39	12976			*						
Classification	9	2	4									

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

As with Cluster no. 18, in Cluster no. 22 a very specific term is used, namely “granularities”, which is not often applied in other contexts (Table 51). Also, it could not establish in the overall dataset over time. Probably, the term became redundant with the in tendency increasing spread of terms like “image processing” (not included in the table). As this term started its rise in 2002, its usage might again not have been common enough for the authors in this cluster to pick it up. Other terms introduced in the term vector are possible methods (“ant colonies” in genetic algorithms) or applications in biomedicine (“cancer”, “HPV”).

Table 52: Keyword ranking and distribution for Cluster no. 42 ("New methods for optimization in clustering").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Evolution	1	2	51						*			
Propositional satisfiability	2	96	2636			*					*	
Binate covering problem	3	51	5555			*						
Numerical optimization	4	48	2522			*			*	*		***
Genetic algorithm (GA) [99]	5	46	33		*			*				
Stochastic algorithms	6	41	6100	*		*			*			
Global optimization	7	32	351		*				*	*		
Greater generosity	8	26	2176	*		*						
Prisoners-dilemma	9	92	2161	*	*				*			*
Branch-and-bound	10	85	2019	*		*		*				*

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

The Cluster no. 42 depicted in Table 52 seems like a very diverse aggregation of terms and documents. However, the backbone of mathematical optimization was for most documents identifiable. Multiple approaches from Mathematics and Game Theory seem to be deployed. The terms "numerical optimization" and "global optimization" had a surge in publication numbers 3 to 6 years later. This could indicate the role these methods played in the emerging topic and how they spread through Artificial Intelligence afterwards.

99 Results for general dataset in ranking and term usage are for "Genetic Algorithm".

Table 53: Keyword ranking and distribution for Cluster no. 45 ("Constraint representation for pattern recognition").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Computational model	1	21	788			**			**		*	
Robotic	2	4	12318			*						*
Real-time	3	6	268	*	**				**			
Algorithms	4	1	5						*			
Eye-movements	5	8	529			*					*	**
Implementation	6	13	220			*			*			
Syntactic ambiguity resolution	7	42	3522			*						
Reference resolution	8	26	2412			*				*		
Disturbed architecture	9	31	5849			*						
Incremental processing	10	32	5826			*					*	

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

From the keywords of Cluster no. 45, it can be deduced that it deals with audio as well as visual patterns that can be used for robotic movements (Table 53). Real-time events have to be split up to single actions that are interpreted. The "syntactic ambiguity resolution", a problem that plays only an inferior role in the remainder of Artificial Intelligence, is of particular interest for the speech recognition. "Eye movement" was another term that can be used for human-computer interaction in particular and gained popularity in 2004 and then in later years again. Thus, this cluster provides yet another example for the combination of topic-specific and ambiguous terms for the definition of a new topic. The older terms are necessary to relate the new topic with their foundations, while only the new terms can enable the efficient communication about new concepts and problems.

Table 54: Keyword ranking and distribution for Cluster no. 46 ("Object delimitation in (moving or static) images").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Primary visual-cortex	1	13	117			*			**			*
Motion	2	50	35						*			
Neurons	3	27	19			*			**			
Dynamics	4	57	21			*			*			
Recognition	5	9	6						*			
Visual-cortex	6	39	75					*				
Spiking neurons	7	94	288			*			*			
Robustness	8	672	517			*	**		**			
Correlation	9	72	667			*		*	*		**	
Restoration	10	8	130		*			*				

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

The term vector for Cluster no. 46 is a mixture of its methods and applications (Table 54). The tight relationship with its biological roots becomes apparent in the top ranked term, which refers to the visual processing for humans. The terms "motion" and "dynamics" both refer to the tracking of moving objects. Both concepts show another increase in frequency in 2007. Again, this cluster shows that a combination of existing methods, also from different scientific fields, is used to built a topic-specific vocabulary.

Table 55: Keyword ranking and distribution for Cluster no. 81 ("Converting vague descriptions to specific ones").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Model	1	50	2									
Context	2	6	211			*			*	*		
Preferred default theories	3	38	4318			*						
Planning with preferences	4	37	4603			*						
Layered neural networks	5	34	5407			*						
Frequent itemset mining	6	32	5454			*		*		*		*
Convertible constraint	7	27	5915			*						
Answer set planning	8	22	2257			*						
Algebraic analysis	9	19	8271		*					*		
Logic programs	10	1	286	*				*		*		

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

Cluster no. 81 combines many rather unpopular methods from Artificial Intelligence for the constraint conversion (Table 55). In particular of interest for the analysis presented here is the term “context”, which also had a rise in publication numbers in 2004. It shows that the meaning of context in general for NLP increased. Such a development might have triggered the emergence of Cluster no. 81 and similar topics. For the remainder, the majority of the terms used here are very specific and uncommon for Artificial Intelligence.

Table 56: Keyword ranking and distribution for Cluster no. 86 (“Stability of fuzzy systems”).

Keyword	**Term rank in…**			**Term usage in whole dataset in the year…**								
	ET cand.	**cluster**	**dataset**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Systems	1	2	1									
Neural-networks[100]	2	6	22		**	*						
Design	3	1	10									
Models	4	5	11						*			
T- and S-norms	5	203	10116			*						
Dynamic-systems	6	17	430			*			*			
Decision-support	7	211	758									
Similarity index	8	222	9011				*		*		*	
System stability	9	223	8945			*						
Controller-design	10	26	2133	*		*				*		*

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to the previous year).

For Cluster no. 86, in the majority ambiguous terms can be found – with two very specific and well-describing exceptions (Table 56): the terms “dynamic-systems” and “system stability” are very representative for the topic of stability of fuzzy systems (if the term fuzzy is not used itself, as it seems to be the case here). Dynamic systems were also a trending topic in 2004 in Artificial Intelligence. This might on the one hand promoted the usage of the term in the keyword set but also on the other hand led to the emergence of this topic itself.

100 Results in ranking and term usage are given for “Neural-Network” as this was the most common spelling of the term. Other variants of spelling (e.g. “Neural Networks”) were omitted for this analysis.

Table 57: Keyword ranking and distribution for Cluster no. 99 ("Collaborative retrieval").

Keyword	Term rank in…			Term usage in whole dataset in the year…								
	ET cand.	cluster	dataset	2002	2003	2004	2005	2006	2007	2008	2009	2010
Web search	1	40	816		*	**		*	*			
Personalized information retrieval	2	29	3097			*						
Information agents	3	7	8025			*			*			
English	4	35	1422			**					*	

Source: WoS, own calculations and illustrations

Notes: The asterisks denote the relative annual growth in term usage in the dataset: *: growth, i.e. relative term usage is higher than in previous year, **: growth, so that usage at least doubles (with regard to the previous year), ***: growth, so that usage at least quintuples (with regard to previous year).

As already explained above (Section 10.2.2), new methods for distributed retrieval came into existence in the last decade. Also, the term "web search" trended in 2004, which might have resulted in a need-driven topic emergence resulting in Cluster no. 99 (Table 57). The term "information agents" was introduced in the context of the combination of multiple retrieval systems. Given the rank in the overall dataset, the term is very specific for the problem in this cluster.

Regarding the findings for these ten exemplary clusters, it can be concluded that there are indeed different concepts of term usage present when a topic emerges:

- Already existing, widely spread terms are used in the context to position the topic in relation to well-known ideas, methods and problems
- New, very specific terms are introduced that summarize the main method or problem of the topic. In most of the examples, these terms were nearly self-explanatory, as they are supposed to deliver the important information "at a glance".
- Terms which are used with an increasing frequency in that year are used to align the topic to current events. Such a trend might be evoked by the introduction of a new method that requires new findings or that can be applied in various contexts.

Main findings in relation to this thesis are that new topics need some time to align their term usage. A topic, as a concept represented by various documents (as described in Section 2.1), is portrayed with a multitude of terms in the beginning. Before the topic is properly defined, coordination about the vocabulary is hampered. Thus, terms are used that might be more specific but also not necessarily common knowledge. This was for instance shown by the absence of the term "fuzzy" in Cluster no. 86. Quite contrary, new terms might be introduced, that later spread in the scientific field. Again, a term representing a method or problem might thus have become applicable for various contexts.

There were no indications that the approach based on the detection of emerging topics was suffering from its heavy reliance on term usage. The clusters and also the emerging topic candidates seemed plausible. The clustered documents shared common concepts which were in turn expressed via the usage of shared keywords.

10.4 Summary

This chapter provided an exemplary analysis of the term usage in emerging topics. The topics were selected with the help of the approach presented in this thesis. All results presented in this chapter suggest that the approach was able to detect emerging topics. However, there can be no statements made for possible missed topics. Also, slight variations in the settings of the approach were necessary to enable its application on the dataset.

Regarding the findings for the term usage, in most cases the term vector of the emerging topic candidates and the overall cluster and dataset differed. In particular, some examples could be found for which otherwise rarely used terms were introduced. In some of these cases, these terms were later better disseminated in the dataset which might hint at the kick-off of a trend. Particular examples are:

- "information agents" for the distributed retrieval
- "qualitative spatial representation" for data analysis for prediction
- "syntactic ambiguity resolution" for pattern recognition in spoken text

These terms did not "solidify" in the scientific landscape. Rather, their usage was specific for the respective topics. They show that the necessary vocabulary was not available at that point in time. To compensate for that, new terms are indeed introduced to cover for the loss of such terms.[101]

In other cases, already existing terms are applied or even combined in an emerging topic, so that they can be used to describe new concepts. The term vectors suggest that already existing vocabulary is in particular adapted when a need-driven topic emerges. In this case, the application, which makes specific tasks and their optimization necessary, is already established. Alternatively, novel methods are introduced, which is represented via new terms for the topic.

The lively interplay between Artificial Intelligence and Biology or Biomedicine, which was already mentioned in Chapter 7, again became apparent in the term vectors that covered respective terms. In particular, biological concepts either introduced the background for a topic (e.g. "primary visual cortex") or its application area (e.g. "gene selection").

Limitations of this study are that keywords had to be used, as they offered more condensed information on the topic specific terms than a full list of terms from the abstracts. Also, the term vectors of the topic model built by the approach were hard to interpret after stemming etc. and with the "contamination" with author names. Because no stemming was performed on the keywords, the actual dissemination could be higher for single terms. Actually, the best solution would have been to implement a semantic analysis of the terms to not only find variations in writings but also synonyms or related terms. However, this as well would have gone beyond the scope of this analysis.

Another point worth mentioning is that the term analysis was only possible on the level of clusters. For that, the emerging topic candidates of a cluster were used to build the "emerging core" of a topic. This stands in contrast to the final approach, which worked on the document level and was detached from

---

101 Note – as was mentioned before – that these terms are particularly figurative, so that their meaning is easy to understand also by non-experts.

the cluster aggregation in the mid-step. Yet, otherwise the structured analysis of introduced vocabulary would have been too dependent on singular choices. On a similar notion, a comparison of actual term combinations could have been analyzed, so that the interactions between terms could have been studied. Again, this has to be postponed for future work, as the scope in the context of this thesis was limited.

The clusters were selected based on their suitability for the contextual and term-based analysis. A bias could have been introduced by this selection in the favour of the approach. However, a document-based analysis did not indicate this. 7 out of the 28 clusters contained documents for which in this more superficial analysis only loose connections could be found. The reader is referred to Chapter 9 for a more substantial proof that the approach is overall working well.

11 Conclusions

This chapter reviews the results of this thesis and discusses the applicability of the developed system with regard to the problem statements:

- Which factors help to differentiate between emerging and established topics (Chapters 6, 7 and 8 and Section 9.4)?
- Is the approach able to identify new emerging topics (Section 9.5)?
- What does the distribution of terms look like for such topics (Chapter 10)?

During this thesis, a closer look was taken at the "hostile environment" in which new topics are born, and how they emerge and develop in it. In particular, internal (Sections 5.2 and 9.3) and external (Chapter 6) features were analyzed to implement a system for the identification of emerging topics on the one hand (Chapter 9), but also to provide a profound understanding of the development of terms during the emergence process on the other hand (Chapter 10). In addition, the role of interdisciplinarity was studied as an indicator for the novelty of emerging topics (Chapter 7). Furthermore, the widespread usage of citation analysis to identify research fronts at the cutting edge was dissected (Section 2.2 and Chapter 8).

The insights gained during the course of this thesis are presented in the next section, which is followed by a discussion of possible extensions to the system and other future applications. Finally, some unanswered questions are presented, which could not be handled within the limited scope of this thesis.

11.1 Synopsis of Results

The goal of this thesis was to develop a system that enables the semi-automatic detection of emerging topics in scientific databases. The system consists of two parts that act as topic filters: The first part, the adapted LDA, uses textual features to create topic models to identify newly introduced topic concepts. The second part uses bibliometric features to look for documents deviating from the "publishing norm". In the course of this thesis, the concept of internal features (those that can be influenced by the authors of a document and that were used in the LDA approach) and external features (those that are influenced by the environment of the publication) was developed. The usage of the latter in the approach was based on findings in a regression analysis (Chapter 6) and a rule derivation (Section 9.4).

Before restricting the approach to a single set of indicators, the two most widely used indicators of a developing emerging topic were analyzed: Interdisciplinarity (Chapter 7) and citations (Chapter 8). As already argued in Section 2.1, bringing together and combining the knowledge, insights, ideas and methods from different research fields is often the departure point for new topics. Therefore, in Chapter 7, the connection between the interdisciplinarity of a topic and its novelty was analyzed. Different approaches to measure the interdisciplinarity of a topic were explored. A high share of innovative work was identified in topics combining knowledge from different disciplines. However, this turned out to be slightly misleading as – of course – many topics were merely using concepts without creating new ones. Despite this, the number of fields in the references (and of the journal, as shown separately) did prove to be an effective feature and was applied in the approach developed here.

The second indicator, citation analysis, is a widely used method for identifying topic clusters or emerging research fronts. Besides the fact that the necessary time lag impedes the usage of citations, it was also assumed that they are an untrustworthy source for innovative research – a notion that could be corroborated in Chapter 8. On the contrary, the research in Chapter 8 showed that the publications in new topics have lower citation rates and experience delays in the first citations and the maximum number of citations. Despite its limitations (in particular with regard to the dataset used), the study confirmed the need for alternative methods to identify emerging topics and for rethinking the citation process.[102]

The LDA adaptation in this thesis was described in Section 5.2 and the respective parameter estimation was realized in Chapter 9. This part of the approach used the internal features of documents to cluster and compare them. All features concerning the document itself were labelled internal features. These included the wording and references of a publication, but also – as a result of the parameter setting – the author names. LDA was extended to make use not only of the terms but also the references and the author names (with different weights) in order to build the document clusters. Based on the instance probabilities (referring to the usage of terms, references and authors), the similarities between clusters could be calculated as well. The so established connections helped to identify predecessors of clusters and, more importantly, clusters without predecessors.

In the LDA approach, the textual components – the abstract, author names and references – were used to deduce the topic model for the document set. Based on the topic model, documents sharing the same topic were clustered. The clusters in the dataset were compared with clusters formed for the previous year, in this case the year 2006. Those clusters that showed no similarity to cluster topics in the previous year were marked as emerging topics. This approach proved to work well on the Training Set. In most cases, the result set was smaller and showed higher Precision than the initial set. Only in Molecular Biology & Genetics was the Precision worse in the LDA result set, and in Pharmacology & Toxicology, the set size decreased while the Precision was unchanged.

To improve the results of the LDA approach even further, the so-called external features were then used to deduce rules for emerging topics. External features are the characteristics of the journals, authors and references that could be calculated in the WoS. Rules were established for individual disciplines. However, they suffered from the same problem as the regression model, in that the whole subject of investigation in this thesis is characterized by a small case number. Thus, the generalizability of the observations in this thesis is always limited. Nonetheless, it could be shown that emerging topics in the individual disciplines are characterized by certain features. Most pronounced, in Engineering, papers in emerging topics are published in journals with a lower article count and a higher age and with the involvement of more authors than those in established topics. In Medicine, collaboration seems to be hampered as the number of authors was even lower for emerging topics. These and other features were used for the rules, which were derived from a resampled document set that suppressed the differences in frequency for established and emerging topics. For most disciplines, the number of fields, the age of the journal as well as the references' age, the number of countries and authors were used in the

---

102 Such rethinking has already been demanded multiple times by others, see e.g. Opthof (1997).

rules. It could be shown that rules can be established for those disciplines in which the number of countries and authors is lower for new documents. There are different trends for the disciplines concerning the age of references and the age and fields of the journal.

The Training Set results legitimated the combination of the two approaches. Thus, the combined approach that uses both the internal and external features of a publication to assess its novelty was tested.

Since the rules differed for the disciplines, a mapping between the disciplines in the Training Set and those in the Test Set had to be found. In the end, the rules of the disciplines Pharmacology & Toxicology, Computer Science and Engineering were tested on four new disciplines: Chemistry, Mathematics, Space Science and Material Science.

The best results in terms of Precision were achieved in Chemistry and Materials Science, with values of 50% and 100%, respectively. Mathematics performed the worst, but further investigations showed that the Recall in this particular result set accounted for 100%. In other cases, documents were lost because the values or thresholds for the disciplines seemed to differ slightly. For instance, in Space Science the number of countries was higher than in Engineering in general.

In all disciplines, the manual effort could be reduced so that only 0.4% to 30% of the initial document set needed to be inspected. In combination with the high Precision values, the results seem promising for future applications. Nonetheless, for better transferability, a detailed comparison between the feature values of different disciplines should be performed.

All in all, it can thus be concluded – based on the results for the Test Set – that the approach was indeed able to

a) find emerging topics in a scientific dataset,
b) replicate the findings of other approaches that relied on citation analysis, and
c) sort out the majority of documents belonging to established topics in the dataset.

This answers the research question about whether it is possible to implement a system for the detection of emerging topics based solely on the features available in real-time. As explained above, different features were tested while developing the system, so that useful indicators of emerging topics were identified. As an additional assessment of the system, Chapter 10 showed – with the help of another testing environment – the usage of terms during the emergence of a topic. The results suggest that the documents in an emerging topic tend to introduce new and very specific terms which spread neither vertically (i.e. across topic boundaries) nor horizontally (i.e. over time).

11.2 Outlook

The approach presented in this thesis could be extended in various ways. Most prominently, features that are not restricted to bibliometric data could be used. The system could also be applied to non-scientific text databases in which the respective metadata could be used analogously.

The thesis concentrated on bibliometric indicators apart from citations. Of course, other information could be used which is not derivable from the bibliometric database used in this thesis. For instance, the affiliation of the authors to institutions and projects was not applied, but could provide more insights and indications of the possibilities and limitations of a topic with regard to dispersion and collaboration. The transfer to a more general context could also provide new indicator options. Altmetrics[103] are currently being established in the bibliometric context. They cover alternative measures of impact like the number of downloads or accesses to an online publication and the number of references to it in Social Media, e.g. Twitter. In this way, Altmetrics could introduce new features in a real-time environment that are similar but – in a sense – more transparent and honest than citations.

The work in this thesis on the role of interdisciplinarity in the emergence process of new scientific topics was limited in respect to the classification scheme. A more in-depth analysis would be necessary to decide whether the results are independent of the classification system. The same criticism applies to the features in the emerging topic detection system, which use the classification system in the background – namely, the number of fields for the journals and the references.

Given the limited data, it was not possible to analyze the role played by the experience and status of the authors in the emergence and dissemination of a topic. Furthermore, full-texts were not available for the document set, so that all analyses had to be conducted on the basis of titles, abstracts and keywords. More precise results would certainly be possible with a dataset providing full texts.

Overall, the approach presented in this thesis represents a first step towards a system to detect emerging topics at an early stage. The focus here was clearly on the applicability of the system in real-time and correspondingly suitable features were selected and assessed. The evaluation results proved to be promising in that the system succeeded in detecting most topics. The transfer of the implications obtained here to similar systems in other contexts is highly recommended, especially using internal and external features to detect not only emerging topics but also impediments to the advancement of scientific frontiers.

---

103 For a definition, see http://altmetrics.org/manifesto/, last accessed on 2014/03/04.

IV Appendix

List of Variables, Parameters and Features

Remark on the notation used in this thesis: For better readability, the capital letters refer to sets of instances themselves as well as to their size. For instance, *M* denotes the set of documents as an aggregation of various instances, as well as the number of documents covered in the set, which would be correctly denoted as |*M*|. For conciseness, formulars by other authors were adapted to use the same variable names and notations as listed below whenever possible.

Context	Abbreviation	Meaning
Bibliometrics	h	The size of the h-core, i.e. the maximum number of publications *h* for a single author that also received at least *h* citations each
General	a	One single author
	d	One single document
	F	(Number of)Features
	f	One single feature
	J	Values of the class feature
	j	One single value of the class feature
	K	Number of topics/clusters generated by an approach
	k	One single topic/cluster
	M	Number of documents/instances
	M_n	Number of “new“” documents
	M_o	Number of “old” documents
	p	probability
	T	Number of actual topics/clusters (in a Gold Standard etc.)
	t	One single genuine topic/cluster, also used for thresholds
	V	Number of terms
	w	One single term (also used relative to its position *n* in a document *d* as *w*(*d*,*n*)), also weight in a function
	X	Values of a feature
	x	One single value of a feature
	Y	A time span, e.g. *Y* years
	y	One single year/also goal value in a regression
LDA	F_t	Textual features used in LDA
	i	Number of iterations of an approach, number of textual input sources for the LDA approach
	n	Average number of documents per topic/cluster or number of documents for a single topic/cluster, also position of a term in a document
	nd	Number of occurrences of terms for a topic in a document

Context	Abbreviation	Meaning
	nr	Number of assignments of a reference to a topic
	nw	Number of assignments of a word to a topic
	R	Number of references
	r	One single reference
	rd	Number of occurrences of references for a topic in a document
	t_w	Threshold for “common words”
	V	Number of terms
	w	One single term (also used relative to its position *n* in a document *d* as *w*(*d*,*n*)), also weight in a function
	w_t	Weight of the textual features
	z	Term-topic assignment, in particular *z(d,n)*, i.e. the topic for the word *w(d,n)*
	z'	Reference-topic assignment
	α	Primer for θ
	β	Parameter for φ
	γ	Parameter for T
	θ	Dirichlet distribution for the per-document topic proportion
	T	Multinomial distribution for the per-topic reference distribution
	φ	Multinomial distribution for the per-topic word distribution
Machine Learning	F_β	F-Measure with parameter β
	H	Entropy
	IG	Information Gain
	P	Precision
	R	Recall
	β	weight for F-Measure (see above)
Regression	cit0,..., cit5	The number of citations in a year 0,..., 5 years after publication of a specific publication
	citTotal	The total number of citations that a publication received in a specific observation period
	firstyear	The year in which a publication is cited for the first time
	maxyear	The year in which a publication reaches its maximum annual citation rate
Regression/Rules	Age of ref.	The average age of the references of a publication; the age is calculated as the time span in years between the current year and the publication year of the respective reference
	Fields of ref.	The number of distinct fields to which the references of a publication are assigned

Context	Abbreviation	Meaning
	JIF	Journal Impact Factor, for definition see Section 3.2.3. The value relates to the journal of the respective publication
	Journal age	The time span between the current year and the year of the first coverage of a journal in the database. The value relates to the journal of the respective publication
	Journal fields	The number of distinct fields to which a journal is assigned in the WoS source classification system. The value relates to the journal of the respective publication
	Journal size	The number of articles a journal published per year. The value relates to the journal of the respective publication
	newTopic	Variable used (especially for regressions) to codify whether a publication belongs to an emerging topic (1) or not (0)
	Nr. of author countries	The number of distinct countries from which the authors are
	Nr. of authors	The number of authors of a publication
Similarity calculation	t_c	Threshold for connections
	w_r	Weight of the references in the similarity calculation of two topics
	Y	The time span in years used in the calculation of connections between the topic clusters

Full List of Content

List of Tables

List of Figures

V Publication bibliography

Oxford dictionary of English (2010). 2nd ed., rev. Oxford: Oxford University Press.

Honorary Authorship (2012). In *Science* 337 (6098), p. 1453.

Abernathy, William J.; Utterback, James M. (1978): Patterns of Industrial Innovation. In *Technology Review* 80 (7), pp. 40–47.

Abernethy, Bruce; Sparrow, W. A. (1992): The Rise and Fall of Dominant Paradigms in Motor Behaviour Research. In Jeffery J. Summers (Ed.): Approaches to the study of motor control and learning, pp. 3–45.

Abraham, P. (2000): Duplicate and salami publications. In *Journal of Postgraduate Medicine* 46, pp. 67–69.

Adams, Jonathan (2005): Early citation counts correlate with accumulated impact. In *Scientometrics* 63 (3), pp. 567–581.

Adams, K. (2006): The sources of innovation and creativity. Paper commissioned by the national center on education and the economy for the new commission on the skills of the American workforce. Washington, DC: National Center on Education and the Economy.

Ahlgren, Per; Colliander, Cristian (2009): Document-document similarity approaches and science mapping: Experimental comparison of five approaches. In *Journal of Informetrics* 3 (1), pp. 49–63. DOI: 10.1016/j.joi.2008.11.003.

Aksnes, Dag W. (2003): A macro study of self-citation. In *Scientometrics* 56 (2), pp. 235–246.

Albright, Kendra (2010): Multidisciplinarity in Information Behavior: Expanding Boundaries or Fragmentation of the Field? In *Libri* 60 (2). DOI: 10.1515/libr.2010.009.

Alexander, David (2012): Our starting point. In *International Journal of Disaster Risk Reduction* 1, pp. 1–4. DOI: 10.1016/j.ijdrr.2012.06.002.

Alonso, S.; Cabrerizo, F.J; Herrera-Viedma, E.; Herrera, F. (2009): h-Index: A review focused in its variants, computation and standardization for different scientific fields. In *Journal of Informetrics* 3 (4), pp. 273–289. DOI: 10.1016/j.joi.2009.04.001.

Alvargonzález, David (2011): Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences. In *International Studies in the Philosophy of Science* 25 (4), pp. 387–403. DOI: 10.1080/02698595.2011.623366.

Amat, C. B.; Yegros Yegros, A. (2009): Median age difference of references as indicator of information update of research groups: A case study in Spanish food research. In *Scientometrics* 78 (3), pp. 447–465. DOI: 10.1007/s11192-007-1993-4.

Anholt, R. M.; Stephen, C.; Copes, R. (2012): Strategies for Collaboration in the Interdisciplinary Field of Emerging Zoonotic Diseases. In *Zoonoses and Public Health* 59 (4), pp. 229–240. DOI: 10.1111/j.1863-2378.2011.01449.x.

Archambault, Eric; Beauchesne, Olivier H.; Caruso, Julie (2011): Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology. In *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics*, pp. 66–77.

Arrow, Kenneth J.; Bernheim, B. Douglas; Feldstein, Martin S.; McFadden, Daniel L.; Poterba, James M.; Solow, Robert M. (2011): 100 Years of the American Economic Review. The Top 20 Articles. In *American Economic Review* 101 (1), pp. 1–8. DOI: 10.1257/aer.101.1.1.

Atkins, Peter W. (2003): Galileo's finger. The ten great ideas of science. Oxford: Oxford University Press.

Baeza-Yates, R.; Ribeiro-Neto, Berthier (1999): Modern information retrieval. New York, Harlow, England: ACM Press; Addison-Wesley.

Baral, C.; Zhao, J. (2008): Non-monotonic temporal logics that facilitate elaboration tolerant revision of goals. In *Proceedings of the national conference on artificial intelligence*, pp. 406–411.

Barber, B. (1961): Resistance by Scientists to Scientific Discovery. In *Science* 1 (September), pp. 596–602.

Bar-Ilan, Judit (2008): Informetrics at the beginning of the 21st century—A review. In *Journal of Informetrics* 2 (1), pp. 1–52. DOI: 10.1016/j.joi.2007.11.001.

Bar-Ilan, Judit (2010): Web of Science with the Conference Proceedings Citation Indexes: the case of computer science. In *Scientometrics* 83 (3), pp. 809–824. DOI: 10.1007/s11192-009-0145-4.

Baron, Jonathan (2000): Thinking and deciding. 3rd ed. Cambridge, UK, New York: Cambridge University Press.

Benos, D. J.; Bashari, E.; Chaves, J. M.; Gaggar, A.; Kapoor, N.; LaFrance, M. et al. (2007): The ups and downs of peer review. In *American Journal of Physics: Advances in Physiology Education* 31 (2), pp. 145–152. DOI: 10.1152/advan.00104.2006.

Blei, David M. (2012): Probabilistic topic models. In *Communications of the ACM* 55 (4), p. 77. DOI: 10.1145/2133806.2133826.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael J. (2003): Latent Dirichlet Allocation. In *Journal of Machine Learning Research* (3), pp. 993–1022.

Blum, E. K.; Lototsky, Sergey V. (2006): Mathematics of physics and engineering. Singapore: World Scientific Publishing Co. Pte. Ltd.

Boden, Margaret A. (1994): What Is Creativity? In Margaret A. Boden (Ed.): Dimensions of creativity. Cambridge (Mass.), London: MIT Press.

Bohlen und Halbach, Oliver von (2011): How to judge a book by its cover? How useful are bibliometric indices for the evaluation of “scientific quality” or “scientific productivity”? In *Annals of Anatomy - Anatomischer Anzeiger* 193 (3), pp. 191–196. DOI: 10.1016/j.aanat.2011.03.011.

Borgman, Christine L. (1999): What are digital libraries? Competing visions. In *Information Processing & Management* 35 (3), pp. 227–243. DOI: 10.1016/S0306-4573(98)00059-4.

Borgman, Christine L. (2007): Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge (Mass.): MIT Press.

Borgman, Christine L.; Rice, Ronald E. (1992): The convergence of information science and communication: A bibliometric analysis. In *Journal of the American Society for Information Science* 43 (6), pp. 397–411.

Börner, Katy; Chen, Chaomei; Boyak, Kevin (2003): Visualizing Knowledge Domains. In Blaise Cronin (Ed.): Annual Review of Information Science & Technology, vol. 37, pp. 179–255.

Bornmann, Lutz (2011): Mimicry in science? In *Scientometrics* 86 (1), pp. 173–177. DOI: 10.1007/s11192-010-0222-8.

Bornmann, Lutz; Daniel, Hans-Dieter (2008): What do citation counts measure? A review of studies on citing behavior. In *Journal of Documentation* 64 (1), pp. 45–80. DOI: 10.1108/00220410810844150.

Bornmann, Lutz; Moya Anegón, Félix de; Leydesdorff, Loet; Valdes-Sosa, Pedro Antonio (2010): Do Scientific Advancements Lean on the Shoulders of Giants? A Bibliometric Investigation of the Ortega Hypothesis. In *PLoS ONE* 5 (10), pp. e13327. DOI: 10.1371/journal.pone.0013327.

Bornmann, Lutz; Mutz, Rüdiger; Daniel, Hans-Dieter (2008): Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. In *Journal of the American Society for Information Science and Technology* 59 (5), pp. 830–837. DOI: 10.1002/asi.20806.

Bornmann, Lutz; Schier, Hermann; Marx, Werner; Daniel, Hans-Dieter (2012): What factors determine citation counts of publications in chemistry besides their quality? In *Journal of Informetrics* 6 (1), pp. 11–18. DOI: 10.1016/j.joi.2011.08.004.

Boyack, K.W; Klavans, R.; Patek, M.; Yoon, P.; Lyle, H.U (2013): An Indicator of Translational Capacity of Biomedical Researchers. In *Proceedings of the 18th International Conference on Science and Technology Indicators, Berlin*, pp. 52–61.

Boyack, Kevin W.; Klavans, Richard (2010): Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? In *Journal of the American Society for Information Science and Technology* 61 (12), pp. 2389–2404.

Boyack, Kevin W.; Newman, David; Duhon, Russell J.; Klavans, Richard; Patek, Michael; Biberstine, Joseph R. et al. (2011): Clustering More than Two Million Biomedical Publications: Comparing the

Accuracies of Nine Text-Based Similarity Approaches. In *PLoS ONE* 6 (3), pp. e18029. DOI: 10.1371/journal.pone.0018029.

Boyce, Bert; Kraft, D.H (1985): Principles and Theories in Information Science. In *Annual Review of Information Science and Technology* 20, pp. 153–178.

Braam, Robert R.; Moed, Henk F.; van Raan, Anthony F. J. (1991): Mapping of science by combined co-citation and word analysis. I. Structural aspects. In *Journal of the American Society for Information Science and Technology* 42 (4), pp. 233–251. Available online at http://www.cwts.nl/TvR/documents/AvR-CoCit-Word-I.pdf.

Braun, Tibor; Glänzel, Wolfgang; Schubert, András (2010): On Sleeping Beauties, Princes and other tales of citation distributions … In *Research Evaluation* 19 (3), pp. 195–202. DOI: 10.3152/095820210X514210;

Broadus, R. N. (1987): Toward a definition of “bibliometrics”. In *Scientometrics* 12 (5-6), pp. 373–379. DOI: 10.1007/BF02016680.

Burrell, Quentin L. (2002): Will this paper ever be cited? In *Journal of the American Society for Information Science and Technology* 53 (3), pp. 232–235. DOI: 10.1002/asi.10031.

Cahlik, Tomas (2000): Comparison of the Maps of Science. In *Scientometrics* 49 (3), pp. 373–387.

Callon, M.; Courtial, J. P.; Laville, F. (1991): Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. In *Scientometrics* 22 (1), pp. 155–205.

Callon, M.; Courtial, J.-P; Turner, W. A.; Bauin, S. (1983): From translations to problematic networks: An introduction to co-word analysis. In *Social Science Information* 22 (2), pp. 191–235. DOI: 10.1177/053901883022002003.

Callon, Michel (2000): Analyse des relations stratégiques entre laboratoires universitaires et entreprises. In *Reseaux* 18 (99), pp. 171–217.

Campanario, Juan Miguel (2009): Rejecting and resisting Nobel class discoveries: accounts by Nobel Laureates. In *Scientometrics* 81 (2), pp. 549–565. DOI: 10.1007/s11192-008-2141-5.

Chang, Chia-Lin; McAleer, Michael; Oxley, Les (2010): Great Expectatrics: Great Papers, Great Journals, Great Econometrics. In *Econometric Reviews* (30 (6)), pp. 583–619.

Chang, Chia-Lin; McAleer, Michael; Oxley, Les (2011): What makes a great journal great in the sciences? Which came first, the chicken or the egg? In *Scientometrics* 87 (1), pp. 17–40. DOI: 10.1007/s11192-010-0335-0.

Cheruvelil, Kendra S.; Soranno, Patricia A.; Weathers, Kathleen C.; Hanson, Paul C.; Goring, Simon J.; Filstrup, Christopher T.; Read, Emily K. (2014): Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills. In *Frontiers in Ecology and the Environment* 12, pp. 31–38.

Chew, Felix S. (1991): Fate of manuscripts rejected for publication in the AJR. In *American Journal of Roentgenology* 156 (March), pp. 627–632.

Chubin, Daryl E. (1976): The Conceptualization of Scientific Specialties. In *The Sociological Quarterly* 17 (4), pp. 448–476.

Cobo, M.J; López-Herrera, A.G; Herrera-Viedma, E.; Herrera, F. (2011): An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. In *Journal of Informetrics* 5 (1), pp. 146–166. DOI: 10.1016/j.joi.2010.10.002.

Costas, Rodrigo; Leeuwen, Thed N.; Raan, Anthony F. J. (2011): The "Mendel syndrome" in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists. In *Scientometrics* 89 (1), pp. 177–205. DOI: 10.1007/s11192-011-0436-4.

Coulter, Neal; Monarch, Ira; Konda, Suresh (1998): Software Engineering as Seen through Its Research Literature: A Study in Co-Word Analysis. In *Journal of the American Society for Information Science and Technology* 49 (13), pp. 1206–1223.

Courtial, J. P. (1994): A coword analysis of scientometrics. In *Scientometrics* 31 (3), pp. 251–260.

Courtial, J. P.; Michelet, B. (1990): A mathematical model of development in a research field. In *Scientometrics* 19 (1-2), pp. 127–141.

Cronin, Blaise (1984): The citation process. London: Taylor Graham.

Cronin, Blaise; Meho, Lokman (2006): Using the h-index to rank influential information scientists. In *Journal of the American Society for Information Science and Technology* 57 (9), pp. 1275–1278. DOI: 10.1002/asi.20354.

Ding, Ying (2011): Community detection: Topological vs. topical. In *Journal of Informetrics* 5 (4), pp. 498–514. DOI: 10.1016/j.joi.2011.02.006.

Dirk, L. (1999): A Measure of Originality: The Elements of Science. In *Social Studies of Science* 29 (5), pp. 765–776. DOI: 10.1177/030631299029005004.

Doreian, Patrick; Fararo, Thomas J. (1985): Structural Equivalence in a Journal Network. In *Journal of the American Society for Information Science* 36 (1), pp. 28–37.

Dorta-González, P.; Dorta-González, M. I. (2013): Impact maturity times and citation time windows: The 2-year maximum journal impact factor. In *Journal of Informetrics* 7 (3), pp. 593–602.

Eckmann, Michael; Rocha, Anderson; Wainer, Jacques (2012): Relationship between high-quality journals and conferences in computer vision. In *Scientometrics* 90 (2), pp. 617–630. DOI: 10.1007/s11192-011-0527-2.

Englert, F.; Brout, R. (1964): Broken Symmetry and the Mass of Gauge Vector Mesons. In *Physical Review Letters* 13 (9), pp. 321–323.

Erosheva, Elena; Fienberg, Stephen; Lafferty, John (2004): Mixed Membership Models of Scientific Publications. In *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl 1), pp. 5229 - 5227.

Fang, Hui (2014): An Explanation of Resisted Discoveries Based on Construal-Level Theory. In *Science and Engineering Ethics* (Epub ahead of print). DOI: 10.1007/s11948-013-9512-x.

Finger, Stanley (2000): Minds behind the brain. A history of the pioneers and their discoveries. Oxford, New York: Oxford University Press.

Franck, G. (1999): Scientific Communication - A Vanity Fair? In *Science* 286 (5437), pp. 53–55.

Fraser, Véronique J.; Martin, James G. (2009): Marketing data: Has the rise of impact factor led to the fall of objective language in the scientific article? In *Respiratory Research* 10 (1), p. 35. DOI: 10.1186/1465-9921-10-35.

Fu, Lawrence D.; Aphinyanaphongs, Yindalon; Wang, Lily; Aliferis, Constantin F. (2011): A comparison of evaluation metrics for biomedical journals, articles, and websites in terms of sensitivity to topic. In *Journal of Biomedical Informatics* 44 (4), pp. 587–594. DOI: 10.1016/j.jbi.2011.03.006.

Garfield, E.; Sher, I. H. (1963): New factors in the evaluation of scientific literature through citation indexing. In *American Documentation* 14 (3), pp. 195–201. DOI: 10.1002/asi.5090140304.

Garfield, Eugene (1970): Would Mendel's Work Have Been Ignored If The Science Citation Index Was Available 100 Years Ago? In *Essays of an Information Scientist* 1 (Current Contents, #2), pp. 69–70.

Garfield, Eugene (1974-76): Is the ratio between number of citations and publications cited a true constant? In *Essays on an Information Scientist* 2 (Current Contents, #6, p.5-7, February 9, 1976), pp. 419–425.

Garfield, Eugene (1979): Is citation analysis a legitimate evaluation tool? In *Scientometrics* 1 (4), pp. 359–375.

Garfield, Eugene (1984): "Science Citation Index" - A New Dimension in Indexing. In *Essays of an Information Scientist* 7, pp. 525–535.

Gigerenzer, Gerd (1994): Where Do New Ideas Come From? In Margaret A. Boden (Ed.): Dimensions of creativity. Cambridge (Mass.), London: MIT Press.

Giuliani, Francesco; Petris, Michele Pio de; Nico, Giovanni (2010): Assessing scientific collaboration through coauthorship and content sharing. In *Scientometrics* 85 (1), pp. 13–28. DOI: 10.1007/s11192-010-0264-y.

Glänzel, W. (2003): Bibliometrics as a research field. A course on theory and application of bibliometric indicators. Course script, Katholieke Universiteit Leuven, Leuven, Belgium.

Glänzel, Wolfgang (2012): Bibliometric methods for detecting and analysing emerging research topics. In *El Profesional de la Informacion* 21 (2), pp. 194–201. DOI: 10.3145/epi.2012.mar.11.

Glänzel, Wolfgang; Garfield, Eugene (2004): The Myth of Delayed Recognition. In *The Scientist* 18 (11), p. 8.

Glänzel, Wolfgang; Schlemmer, Balázs; Thijs, Bart (2003): Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. In *Scientometrics* 58 (3), pp. 571–586.

Glänzel, Wolfgang; Schoepflin, Urs (1999): A bibliometric study of reference literature in the sciences and social sciences. In *Information Processing & Management* 35 (1), pp. 31–44. DOI: 10.1016/S0306-4573(98)00028-4.

Glänzel, Wolfgang; Schubert, András; Thijs, Bart; Debackere, Koenraad (2011): A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. In *Scientometrics* 87 (2), pp. 415–424. DOI: 10.1007/s11192-011-0345-6.

Glänzel, Wolfgang; Thijs, Bart (2004): The influence of author self-citations on bibliometric macro indicators. In *Scientometrics* 59 (3), pp. 281–310.

Greenland, P.; Fontanarosa, P. B. (2012): Ending Honorary Authorship. In *Science* 337 (6098), p. 1019. DOI: 10.1126/science.1224988.

Griffith, B. C.; Small, H. G.; Stonehill, J. A.; Dey, S. (1974): The Structure of Scientific Literatures II: Toward a Macro- and Microstructure for Science. In *Social Studies of Science* 4 (4), pp. 339–365. DOI: 10.1177/030631277400400402.

Griffiths, Thomas L.; Steyvers, Mark (2004): Finding scientific topics. In *PNAS* 101 (suppl. 1), pp. 5228–5235.

Grinnell, Frederick (1987): The scientific attitude. Boulder: Westview Press.

Grün, Bettina; Hornik, Kurt (2011): Topicmodels: An R Package for Fitting Topic Models. In *Journal of Statistical Software* 40 (13), pp. 1–30.

Grupp, Hariolf (1998): Foundations of the Economics of Innovation—Theory, Measurement and Practice: Edward Elgar, Cheltenham.

Guo, Hanning; Weingart, Scott; Börner, Katy (2011): Mixed-indicators model for identifying emerging research areas. In *Scientometrics* 89 (1), pp. 421–435. DOI: 10.1007/s11192-011-0433-7.

Handy, C. (1995): Trust and the virtual organization. In *Harvard Business Review* 73 (3), pp. 40–50.

Harwood, Nigel (2008): Publication outlets and their effect on academic writers' citations. In *Scientometrics* 77 (2), pp. 253–265. DOI: 10.1007/s11192-007-1955-x.

Harzing, Anne-Wil (2013): Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? In *Scientometrics* 93 (1), pp. 23–34.

Havemann, Frank (2009): Einführung in die Bibliometrie. 1. Aufl. Berlin: Gesellschaft für Wissenschaftsforschung, Institut für Bibliotheks- und Informationswissenschaft.

He, Qin (1999): Knowledge Discover through Co-word Analysis. In *Library Trends* 48, pp. 133–159.

Heinrich, Gregor (2008): Parameter Estimation for Text Analysis. Technical Report.

Hewings, Ann; Lillis, Theresa; Vladimirou, Dimitra (2010): Who's citing whose writings? A corpus based study of citations as interpersonal resource in English medium national and English medium international journals. In *Journal of English for Academic Purposes* 9 (2), pp. 102–115. DOI: 10.1016/j.jeap.2010.02.005.

Higgs, Peter W. (1964): Broken Symmetries and the Masses of Gauge Bosons. In *Physical Review Letters* 13 (16), pp. 508–509.

Hirsch, J. E. (2005): An index to quantify an individual's scientific research output. In *PNAS* 102 (46), pp. 16569–16572.

Hirsch, J. E. (2007): Does the h index have predictive power? In *PNAS* 104 (49), pp. 19193–19198.

Hoekman, Jarno; Frenken, Koen; Oort, Frank (2009): The geography of collaborative knowledge production in Europe. In *The Annals of Regional Science* 43 (3), pp. 721–738. DOI: 10.1007/s00168-008-0252-9.

Hönekopp, Johannes; Kleber, Janet (2008): Sometimes the impact factor outshines the H index. In *Retrovirology* 5 (1), p. 88. DOI: 10.1186/1742-4690-5-88.

Hook, E. B. (2002): A background to prematurity and resistance to “discovery”. In Ernest B. Hook (Ed.): Prematurity in scientific discovery. On resistance and neglect. Berkeley: University of California Press, pp. 3–21.

Horlings, Edwin; Gurney, Thomas (2012): Search strategies along the academic lifecycle. In *Scientometrics* 94 (3). DOI: 10.1007/s11192-012-0789-3.

Huang, M.-H; Chang, Y.-W (2011): A study of interdisciplinarity in information science: using direct citation and co-authorship analysis. In *Journal of Information Science* 37 (4), pp. 369–378. DOI: 10.1177/0165551511407141.

Hummon, Norman P.; Doreian, Patrick (1989): Connectivity in a citation network: The development of DNA theory. In *Social Networks* 11 (1), pp. 39–63.

Ibáñez, Alfonso; Larrañaga, Pedro; Bielza, Concha (2011): Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. In *Scientometrics* 89 (2), pp. 523–551. DOI: 10.1007/s11192-011-0486-7.

Iltis, Hugo (1932): Life of Mendel. Translated by Eden Paul, Cedar Paul: W.W. Norton & Company, Incorporated.

Jacso, Peter (2005): As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. In *Current Science* 89 (9), pp. 1537–1547.

Janssens, Frizo; Glänzel, Wolfgang; Moor, Bart (2008): A hybrid mapping of information science. In *Scientometrics* 75 (3), pp. 607–631. DOI: 10.1007/s11192-007-2002-7.

Janssens, Frizo; Leta, Jacqueline; Glänzel, Wolfgang; Moor, Bart de (2006): Towards mapping library and information science. In *Information Processing & Management* 42 (6), pp. 1614–1642. DOI: 10.1016/j.ipm.2006.03.025.

Jin, BiHui; Liang, LiMing; Rousseau, Ronald; Egghe, Leo (2007): The R- and AR-indices: Complementing the h-index. In *Chinese Science Bulletin* 52 (6), pp. 855–863. DOI: 10.1007/s11434-007-0145-9.

Johnson, Steven (2013): Wo gute Ideen herkommen. Eine kurze Geschichte der Innovation. Bad Vilbel: Scoventa.

Johnston, David W.; Piatti, Marco; Torgler, Benno (2013): Citation success over time: theory or empirics? In *Scientometrics* 95 (3), pp. 1023–1029. DOI: 10.1007/s11192-012-0910-7.

Kajikawa, Yuya; Takeda, Yoshiyuki (2009): Citation network analysis of organic LEDs. In *Technological Forecasting and Social Change* 76 (8), pp. 1115–1123. DOI: 10.1016/j.techfore.2009.04.004.

Kehtarnavaz, N.; Sohn, W. (1991): Steering Control of Autonomous Vehicles by Neural Networks. In *American Control Conference, 26-28 June 1991*, pp. 3096–3101.

Kelly, Clint D.; Jennions, Michael D. (2007): H-index: age and sex make it unreliable. In *Nature* 449, p. 403.

Kilwein, J.H (1999): Biases in medical literature. In *Journal of Clinical Pharmacy and Therapeutics* 24 (6), pp. 393–396.

Klavans, Richard; Boyack, Kevin W. (2006a): Identifying a better measure of relatedness for mapping science. In *Journal of the American Society for Information Science and Technology* 57 (2), pp. 251–263. DOI: 10.1002/asi.20274.

Klavans, Richard; Boyack, Kevin W. (2006b): Quantitative evaluation of large maps of science. In *Scientometrics* 68 (3), pp. 475–499.

Klavans, Richard; Boyack, Kevin W. (2009): Toward a consensus map of science. In *Journal of the American Society for Information Science and Technology* 60 (3), pp. 455–476. DOI: 10.1002/asi.20991.

Klavans, Richard; Boyack, Kevin W. (2010): Toward an objective, reliable and accurate method for measuring research leadership. In *Scientometrics* 82 (3), pp. 539–553. DOI: 10.1007/s11192-010-0188-6.

Kohonen, Teuvo (1982): Self-organized formation of topologically correct feature maps. In *Biological Cybernetics* 43 (1), pp. 59–69. DOI: 10.1007/BF00337288.

Kousha, Kayvan; Thelwall, Mike (2012): Motivations for Citing YouTube Videos in the Academic Publications: A Contextual Analysis. In *Proceedings of the STI conference*, pp. 488–497.

Kozak, Marin (2013): Current Science has its 'Sleeping Beauties'. In *Current Science* 104 (9), pp. 1129–1130.

Kuhn, Thomas S. (1973): The Structure of Scientific Revolutions. Second Edition, Enlarged: The University of Chicago Press.

Lancaster, F. W.; Lee, Sun-Yoon Kim; Diluvio, Catalina (1990): Does place of publication influence citation behavior? In *Scientometrics* 19 (3-4), pp. 239–244.

Lee, Hyoung-joo; Lee, Sungjoo; Yoon, Byungun (2011): Technology clustering based on evolutionary patterns: The case of information and communications technologies. In *Technological Forecasting and Social Change* 78 (6), pp. 953–967. DOI: 10.1016/j.techfore.2011.02.002.

Leimu, R.; Koricheva, Julia (2005a): Does Scientific Collaboration Increase the Impact of Ecological Articles? In *BioScience* 55 (5), pp. 438–443.

Leimu, R.; Koricheva, Julia (2005b): What determines the citation frequency of ecological papers? In *Trends in Ecology & Evolution* 20 (1), pp. 28–32. DOI: 10.1016/j.tree.2004.10.010.

Levitt, Jonathan M.; Thelwall, Mike (2008): Is multidisciplinary research more highly cited? A macrolevel study. In *Journal of the American Society for Information Science and Technology* 59 (12), pp. 1973–1984. DOI: 10.1002/asi.20914.

Leydesdorff, Loet (1987): Various Methods for the Mapping of Science. In *Scientometrics* 11 (5-6), pp. 295–324.

Leydesdorff, Loet (1997): Why words and co-words cannot map the development of the sciences. In *Journal of the American Society for Information Science* 48 (5), pp. 418–427.

Leydesdorff, Loet (1998): Theories of Citation? In *Scientometrics* 43 (1), pp. 5–25.

Leydesdorff, Loet (2003): Can networks of journal-journal citations be used as indicators of change in the social sciences? In *Journal of Documentation* 59 (1), pp. 84–104. DOI: 10.1108/00220410310458028.

Leydesdorff, Loet (2007): "Betweenness Centrality" as an Indicator of the "Interdisciplinarity" of Scientific Journals. In *Journal of the American Society for Information Science and Technology* 58 (9), pp. 1303–1319.

Leydesdorff, Loet; Bornmann, Lutz (2011): How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. In *Journal*

of the American Society for Information Science and Technology 62 (2), pp. 217–229. DOI: 10.1002/asi.21450.

Leydesdorff, Loet; Opthof, Tobias (2010): Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. In *Journal of the American Society for Information Science and Technology* 61 (11), pp. 2365–2369.

Leydesdorff, Loet; Rafols, Ismael (2011): The Local Emergence and Global Diffusion of Research Technologies: An Exploration of Patterns of Network Formation. In *Journal of the American Society for Information Science and Technology* 62 (5), pp. 846–860.

Leydesdorff, Loet; Zhou, Ping; Bornmann, Lutz (2013): How can journal impact factors be normalized across fields of science? An assessment in terms of percentile ranks and fractional counts. In *Journal of the American Society for Information Science and Technology* 64 (1), pp. 96–107.

Li, Linjing; Li, Xin; Cheng, Changjian; Chen, Cheng; Ke, Guanyan; Zeng, Daniel Dajun; Scherer, William T. (2010): Research Collaboration and ITS Topic Evolution: 10 Years at T-ITS. In *IEEE Transactions on Intelligent Transportation Systems* 11 (3), pp. 517–523. DOI: 10.1109/TITS.2010.2059070.

Liu, Mengxiong (1993): Progress in documentation the complexities of citation practice: A review of citation studies. In *Journal of Documentation* 49 (4), pp. 370–408. DOI: 10.1108/eb026920.

Long, J. S. (1997): Regression Models for Categorical and Limited Dependent Variables, Advanced Quantitative Techniques in the Social Sciences Series 7: Thousand Oaks, London, New Delhi: SAGE Publications.

Long, J. S.; Freese, J. (2006): Regression Models for Categorical Dependent Variables Using Stata. Second Edition: College Station, Texas: Stata Press.

Lowry, O. H.; Rosebrougi, N. J.; Farr, A. L.; Randal, R. J. (1951): Protein measurement with the Folin phenol reagent. In *Journal of Bioiogical Chemistry* 193, pp. 265–275.

MacRoberts, M. H.; MacRoberts, Barbara R. (1996): Problems of citation analysis. In *Scientometrics* 36 (3), pp. 435–444.

MacRoberts, Michael H.; MacRoberts, Barbara R. (1989): Problems of citation analysis: A critical review. In *Journal of the American Society for Information Science* 40 (5), pp. 342–349.

Malin, Morton V. (1968): The Science Citation Index. A New Concept in Indexing. In *Library Trends* 16 (3), pp. 374–387.

Mallig, Nicolai (2010): A relational database for bibliometric analysis. In *Journal of Informetrics* 4 (4), pp. 564–580. DOI: 10.1016/j.joi.2010.06.007.

Mane, Ketan K.; Börner, Katy (2004): Mapping topics and topic bursts in PNAS. In *Proceedings of the National Academy of Sciences of the United States of America*, 101 (suppl. 1), pp. 5287–5290.

Mann, Gideon S.; Mimno, David; McCallum, Andrew (2006): Bibliometric Impact Measures Leveraging Topic Analysis. In *JCDL '06: Proceedings of the Joint Conference on Digital Libraries*.

Marin, Jean-Michel; Mengersen, Kerrie; Robert, Christian P. (2005): Bayesian Modelling and Inference on Mixtures of Distributions. In C. Rao, D. Dey (Eds.): Handbook of Statistics, vol. 25. New York: Springer-Verlag, pp. 459–507.

McCain, K. W. (1990): Mapping authors in intellectual space: A technical overview. In *Journal of the American Society for Information Science* 41, pp. 433–443.

Meho, Lokman I.; Rogers, Yvonne (2008): Citation counting, citation ranking, and h-index of Human-Computer Interaction Reseachers. A Comparison of Scopus and Web of Science. In *Journal of the American Society for Information Science and Technology* 59 (11), pp. 1711–1726. DOI: 10.1002/asi.20874.

Meho, Lokman I.; Yang, Kiduk (2007): Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar. In *Journal of the American Society for Information Science and Technology* 58 (13), pp. 2105–2125. DOI: 10.1002/asi.20677.

Mendel, G. (1865): Versuche über Pflanzen-Hybriden (Experiments with Plant Hybrids). In *Proceedings of the National History Society of Brunn (Bohemia, now Czech Republic)*.

Merton, Robert K. (1968): The Matthew Effect in Science. In *Science* 159, pp. 56–68.

Michels, Carolin (under review): Varieties in Quantity and Timing of Citations for Topics in Different Development Stages and Disciplines.

Michels, Carolin (2013): The Relationship between a Topic's Interdisciplinarity and its Innovativeness. Poster presentation. In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna; 15-19 July 2013*, pp. 2105–2108.

Michels, Carolin; Fu, Junying (2014): Systematic Analysis of Coverage and Usage of Conference Proceedings in Web of Science. In *Scientometrics* May (online first).

Michels, Carolin; Neuhäusler, Peter (draft): Towards an Early-Stage Identification of Emerging Topics in Science – The Usability of Bibliometric Characteristics.

Michels, Carolin; Rettinger, Achim (2014): Emerging Topics in Science. Subproject in the Kompetenzzentrum Bibliometrie. Fraunhofer ISI. Karlsruhe (Discussion Papers Innovation Systems and Policy Analysis, 42). Available online at http://www.isi.fraunhofer.de/isi-media/docs/p/de/diskpap_innosysteme_policyanalyse/discussionpaper_42_2014.pdf.

Michels, Carolin; Schmoch, Ulrich (2012): The growth of science and database coverage. In *Scientometrics* 93 (3), pp. 831–846.

Michels, Carolin; Schmoch, Ulrich (2014): Impact of bibliometric studies on the publication behaviour of authors. In *Scientometrics* 98 (1), pp. 369–385. DOI: 10.1007/s11192-013-1015-7.

Mina, A.; Ramlogan, R.; Tampubolon, G.; Metcalfe, J. (2007): Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. In *Research Policy* 36 (5), pp. 789–806. DOI: 10.1016/j.respol.2006.12.007.

Mitchell, Tom M. (1997): Machine learning. New York, NY: McGraw-Hill (McGraw-Hill series in Computer Science).

Moed, H. F.; van Leeuwen, T. N. (1995): Improving the accuracy of institute for scientific information's journal impact factors. In *Journal of the American Society for Information Science* 46 (6), pp. 461–467.

Moore, Andrew W. (2003): Information Gain. School of Computer Science, Carnegie Mellon University. Available online at http://www.autonlab.org/tutorials/infogain11.pdf, checked on 12/13/2013.

Morillo, Fernanda; Bordons, Maria; Gomez, Isabel (2001): An approach to interdisciplinarity through bibliometric indicators. In *Scientometrics* 51 (1), pp. 203–222.

Morris, Steven A.; van der Veer Martens, Betsy (2008): Mapping research specialties. In *Annual Review of Information Science and Technology* 42 (1), pp. 213–295. DOI: 10.1002/aris.2008.1440420113.

Muñoz-Leiva, Francisco; Viedma-del-Jesús, María Isabel; Sánchez-Fernández, Juan; López-Herrera, Antonio Gabriel (2012): An application of co-word analysis and bibliometric maps for detecting the most highlighting themes in the consumer behaviour research from a longitudinal perspective. In *Quality & Quantity* 46 (4), pp. 1077–1095. DOI: 10.1007/s11135-011-9565-3.

Nallapati, Ramesh; Ahmed, Amr; Xing, Eric P.; Cohen, William W. (2008): Joint Latent Topic Models for Text and Citations. In *KDD'08*, pp. 542–550.

Narin, Francis; Pinski, Gabriel; Hofer Gee, Helen (1976): Structure of the biomedical literature. In *Journal of the American Society for Information Science* 27 (1), pp. 25–45.

Nemeth, Charlan Jeanne (1995): Dissent as driving cognition, attitudes, and judgments. In *Social Cognition* 13 (3), pp. 273–291.

Newman, M. E. J. (2001): Who is the best connected scientist? A study of scientific coauthorship networks. In *SFI Working Paper* 00-12-64.

Nickerson, Raymond S. (1998): Confirmation Bias: A Ubiquitous Phenomenon in many Guises. In *Review of General Psychology* 2 (2), pp. 175–220.

Noyons, Ed C. M.; Moed, Henk F.; Luwel, Marc (1999): Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. In *Journal of the American Society for Information Science* 50 (2), pp. 115–131.

Noyons, Ed C. M.; Moed, Henk F.; van Raan, Anthony F. J. (1999): Integrating Research Performance Analysis and Science Mapping. In *Scientometrics* 46 (3), pp. 591–604.

O'Brien, R.M (2007): A Caution Regarding Rules of Thumb for Variance Inflation Factors. In *Quality & Quantity* 41, pp. 673–690.

Ogburn, William F.; Thomas, Dorothy (1922): Are Inventions Inevitable? A Note on Social Evolution. In *Political Science Quarterly* 37 (1), pp. 83–98.

Ogden, T. L.; Bartley, D. L. (2008): The Ups and Downs of Journal Impact Factors. In *Annals of Occupational Hygiene* 52 (2), pp. 73–82. DOI: 10.1093/annhyg/men002.

Ohba, Norio; Nakao, Kumiko (2012): Sleeping beauties in ophthalmology. In *Scientometrics* 93 (2), pp. 253–264. DOI: 10.1007/s11192-012-0667-z.

Olson, Gary M.; Olson, Judith S. (2000): Distance matters. In *Human Computer Interaction* 15 (2), pp. 139–178.

O'Luanaigh, Cian (2013): New results indicate that new particle is a Higgs boson. CERN. Available online at http://home.web.cern.ch/about/updates/2013/03/new-results-indicate-new-particle-higgs-boson, updated on 10/7/2013, checked on 11/15/2013.

Opthof, Tobias (1997): Sense and nonsense about the impact factor. In *Cardiovascular Research* 33, pp. 1–7.

Pepe, Alberto; Rodriguez, Marko A. (2010): Collaboration in sensor network research: An in-depth longitudinal analysis of assortative mixing patterns. In *Scientometrics* 84 (3), pp. 687–701. DOI: 10.1007/s11192-009-0147-2.

Perry, M. (2001): Shared trust in small countries: The limits to borrowing models. In *New Economy* 8 (3), pp. 175–177.

Pinski, Gabriel; Narin, Francis (1976): Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. In *Information Processing & Management* 12, pp. 297–312.

Pollman, Thijs (2000): Forgetting and the Ageing of Scientific Publications. In *Scientometrics* 47 (1), pp. 43–54. DOI: 10.1023/A:1005613725039.

Porter, A. L.; Chubin, D. E. (1985): An indicator of cross-disciplinary research. In *Scientometrics* 8, pp. 161–176.

Porter, Alan L.; Rafols, Ismael (2009): Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. In *Scientometrics* 81 (3), pp. 719–745. DOI: 10.1007/s11192-008-2197-2.

Porter, M. F. (1980): An algorithm for suffix stripping. In *Program* 14 (3), pp. 130–137.

Potter, W.G (1981): Bibliometrics. In *Library Trends* 30, pp. 3–172.

Price, Derek de Solla (1965): Networks of Scientific Papers:The pattern of bibliographic references indicates the nature of the scientific research front. In *Science Communication* 149 (3683), pp. 510–515.

Price, Derek de Solla (1966): Science: The Science of Scientists. In *Medical Opinion and Review* 1 (July), pp. 88–97.

Priddy, Kevin L.; Keller, Paul E. (2005): Artificial Neural Networks: An Introduction. SPIE Press (Vol. 86).

Pritchard, Alan (1969): Statistical bibliography or bibliometrics? In *Journal of Documentation* 25 (4), pp. 348–349.

Racherla, Pradeep; Hu, Clark (2010): A social network perspective of tourism research collaborations. In *Annals of Tourism Research* 37 (4), pp. 1012–1034. DOI: 10.1016/j.annals.2010.03.008.

Rafols, Ismael; Leydesdorff, Loet (2009): Content-based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects. In *Journal of the American Society for Information Science and Technology* 60 (9), pp. 1823–1835.

Ragone, Azzurra; Mirylenka, Katsiaryna; Casati, Fabio; Marchese, Maurizio (2013): On peer review in computer science: analysis of its effectiveness and suggestions for improvement. In *Scientometrics* 97 (2), pp. 317–356. DOI: 10.1007/s11192-013-1002-z.

Ray, Joel; Berkwits, Michael; Davidoff, Frank (2000): The fate of manuscripts rejected by a general medical journal. In *The American journal of medicine* 109 (2), pp. 131–135.

Rinia, Ed J.; van Leeuwen, Thed N.; Bruins, Eppo E. W.; van Vuren, Hendrik G.; van Raan, Anthony F. J. (2001): Citation delay in interdisciplinary knowledge exchange. In *Scientometrics* 51 (1), pp. 293–309.

Roberts, Jason (2009): An Author's Guide to Publication Ethics: A Review of Emerging Standards in Biomedical Journals. In *Headache: The Journal of Head and Face Pain* 49 (4), pp. 578–589. DOI: 10.1111/j.1526-4610.2009.01379.x.

Rocco, E. (1998): Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *Proceedings of the CHI'98 Conference on Human Factors in Computing Systems, New York: ACM.*, pp. 496–502.

Roolaht, Tõnu (2012): The Characteristics of Small Country National Innovation Systems. In Elias G. Carayannis, Umas Varblane, Tõnu Roolaht (Eds.): Innovation Systems in Small Catching-Up Economies. New Perspectives on Practice and Policy, vol. 15. New York: Springer, pp. 21–37.

Rosvall, M.; Bergstrom, C. T. (2008): Maps of random walks on complex networks reveal community structure. In *PNAS* 105 (4), pp. 1118–1123.

Rousseau, Ronald (2007): The influence of missing publications on the Hirsch index. In *Journal of Informetrics* 1 (1), pp. 2–7. DOI: 10.1016/j.joi.2006.05.001.

Ruef, Martin (2002): Strong ties, weak ties and islands: structural and cultural predictors of organizational innovation. In *Industrial and Corporate Change* 11 (3), pp. 427–449, checked on 2/11/2014.

Russell, A. Wendy; Wickson, Fern; Carew, Anna L. (2008): Transdisciplinarity: Context, contradictions and capacity. In *Futures* 40 (5), pp. 460–472. DOI: 10.1016/j.futures.2007.10.005.

Rzeszutek, Richard; Androutsos, Dimitrios; Kyan, Matthew (2010): Self-Organizing Maps for Topic Trend Discovery. In *IEEE Signal Process. Lett* 17 (6), pp. 607–610. DOI: 10.1109/LSP.2010.2048940.

Saka, A.; Igami, M.; Kuwahara, T. (2010): Science Map 2008, NISTEP Report No. 139.

Schaffer, Simon (1994): Making Up Discovery. In Margaret A. Boden (Ed.): Dimensions of creativity. Cambridge (Mass.), London: MIT Press.

Schneider, Jesper W. (2012): Testing University Rankings Statistically: Why this Perhaps is not such a Good Idea after All. Some Reflections on Statistical Power, Effect Size, Random Sampling and Imaginary Populations. In *Proceedings of 17th International Conference on Science and Technology Indicators*, pp. 719–732.

Schubert, Torben; Michels, Carolin (2013): Placing articles in the large publisher nations: Is there a "free lunch" in terms of higher impact? In *Journal of the American Society for Information Science and Technology* 64 (3), pp. 596–611. DOI: 10.1002/asi.22759.

Sebastiani, Fabrizio (2002): Machine Learning in Automated Text Categorization. In *ACM Computing Surveys* 34 (1), pp. 1–47.

Seglen, P. O. (1997): Why the impact factor of journals should not be used for evaluating research. In *British Medical Journal* 314 (7079), p. 497. DOI: 10.1136/bmj.314.7079.497.

Sejnowski, T. J.; Rosenberg, C. R. (1986): NETtalk: a parallel network that learns to read aloud. In *Cognitive Science* 14, pp. 179–211.

Shafique, Muhammad (2013): Thinking inside the box? Intellectual structure of the knowledge base of innovation research (1988-2008). In *Strategic Management Journal* 34 (1), pp. 62–93. DOI: 10.1002/smj.2002.

Sharabchiev, J.T (1989): Cluster analysis of bibliographic references as a scientometric method. In *Scientometrics* 15 (1), pp. 127–137.

Shibata, Naoki; Kajikawa, Yuya; Takeda, Yoshiyuki; Matsushima, Katsumori (2008): Detecting emerging research fronts based on topological measures in citation networks of scientific publications. In *Technovation* 28 (11), pp. 758–775. DOI: 10.1016/j.technovation.2008.03.009.

Shibata, Naoki; Kajikawa, Yuya; Takeda, Yoshiyuki; Matsushima, Katsumori (2009a): Comparative study on methods of detecting research fronts using different types of citation. In *Journal of the American Society for Information Science and Technology* 60 (3), pp. 571–580. DOI: 10.1002/asi.20994.

Shibata, Naoki; Kajikawa, Yuya; Takeda, Yoshiyuki; Sakata, I.; Matsushima, Katsumori (2009b): Early Detection of Innovations from Citation Networks. In *Industrial Engineering and Engineering Management, IEEE International Conference*, pp. 54–58.

Shibata, Naoki; Kajikawa, Yuya; Takeda, Yoshiyuki; Sakata, Ichiro; Matsushima, Katsumori (2009c): Detecting Emerging Research Fronts in Regenerative Medicine by Citation Network Analysis of Scientific Publications. In *PICMET 2009 Proceedings*, August 2-6, Portland, Oregon USA.

Sigelman, Lee (2009): Are Two (or Three or Four … or Nine) Heads Better than One? Collaboration, Multidisciplinarity, and Publishability. In *PS: Political Science & Politics* 42 (03), p. 507. DOI: 10.1017/S1049096509090817.

Sigogneau, Anne (2000): An Analysis of Document Types Published in Journals Related to Physics: Proceeding Papers Recorded in the Science Citation Index Database. In *Scientometrics* 47 (3), pp. 589–604.

Small, H. (1977): A Co-Citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research. In *Social Studies of Science* 7, pp. 139–166.

Small, H.; Griffith, B. C. (1974): The Structure of Scientific Literatures I: Identifying and Graphing Specialties. In *Social Studies of Science* 4 (1), pp. 17–40. DOI: 10.1177/030631277400400102.

Small, Henry (1985): Clustering the science citation index® using co-citations. I. A comparison of methods. In *Scientometrics* 7 (3-6), pp. 391–409.

Small, Henry (2006): Tracking and predicting growth areas in science. In *Scientometrics* 68 (3), pp. 595–610.

Small, Henry (2011): Interpreting maps of science using citation context sentiments: a preliminary investigation. In *Scientometrics* 87 (2), pp. 373–388. DOI: 10.1007/s11192-011-0349-2.

Small, Henry; Boyack, Kevin W.; Klavans, Richard (2014): Identifying emerging topics in science and technology. In *Research Policy* In Press. DOI: 10.1016/j.respol.2014.02.005.

Small, Henry; Garfield, E. (1985): The geography of science: disciplinary and national mappings. In *Journal of Information Science* 11 (4), pp. 147–159.

Small, Henry; Sweeney, E.; Greenlee, E. (1985): Clustering the science citation index using co-citations. II. Mapping science. In *Scientometrics* 8 (5-6), pp. 321–340.

Small, Henry; Upham, Phineas (2009): Citation structure of an emerging research area on the verge of application. In *Scientometrics* 79 (2), pp. 365–375. DOI: 10.1007/s11192-009-0424-0.

Soh, Nerissa; Walter, Garry; Touyz, Stephen; Russell, Janice; Malhi, Gin S.; Hunt, Glenn E. (2012): Food for thought: Comparison of citations received from articles appearing in specialized eating disorder journals versus general psychiatry journals. In *International Journal of Eating Disorders* 45 (8), pp. 990–994. DOI: 10.1002/eat.22036.

Staab, S.; Studer, R.; Schnurr, H.P.; Sure, Y. (2001): Knowledge processes and ontologies. In *Intelligent Systems, IEEE* 16 (1), pp. 26–34.

Statistisches Bundesamt (2013): Bildung und Kultur, Studierende an Hochschulen, Fächersystematik. Wiesbaden (Fachserie 11, Reihe 4.1). Available online at https://www.destatis.de/DE/Methoden/Klassifikationen/BildungKultur/StudentenPruefungsstatistik.pdf.

Stent, G. S.; Hook, E. (2002): Prematurity in scientific discovery. On resistance and neglect. Berkeley, California: University of California Press.

Stent, G.S (1972): Prematurity and Uniqueness in Scientific Discoveries. In *Scientific American* 227 (6), pp. 84–93.

Steyvers, Mark; Griffiths, Thomas (2007): Probabilistic Topic Models. In T. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds.): Latent Semantic Analysis: A Road to Meaning: Laurence Erlbaum.

Stifterverband für die deutsche Wissenschaft (2013): FuE-Datenreport 2013. Available online at http://www.stifterverband.org/publikationen_und_podcasts/wissenschaftsstatistik/fue_datenreport/fue_datenreport_2013_analysen_und_vergleiche.pdf.

Sutton, Richard S.; Barto, Andrew G. (1998): Reinforcement Learning: An Introduction. MIT Press.

Swanson, Don R. (1993): Intervening in the life cycles of scientific knowledge. In *Library Trends* 41 (4), pp. 606–631.

Tagliacozzo, Renata (1977): Self-Citations in Scientific Literature. In *Journal of Documentation* 33 (4), pp. 251–265. DOI: 10.1108/eb026644.

Takahashi, K.; Aw, T.-C.; Koh, D. (1999): An alternative to journal-based impact factors. In *Occupational Medicine* 49, pp. 57–59.

Takeda, Yoshiyuki; Kajikawa, Yuya (2009): Optics: a bibliometric approach to detect emerging research domains and intellectual bases. In *Scientometrics* 78 (3), pp. 543–558. DOI: 10.1007/s11192-007-2012-5.

Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2006): Introduction to Data Mining. Boston: Pearson Education.

The PLoS Medicine Editors (2006): The Impact Factor Game. In *PLoS Medicine* 3 (6), pp. 707f. DOI: 10.1371/journal.pmed.0030291.

Thompson, Dennis F.; Callen, Erin C.; Nahata, Milap C. (2009): New Indices in Scholarship Assessment. In *American Journal of Pharamceutical Education* 73 (6, Article 111).

Thompson Klein, Julie (1996): Crossing boundaries. Knowledge, disciplinarities, and interdisciplinarities. Charlottesville, Va: University Press of Virginia.

Thompson Klein, Julie (2004): Prospects for transdisciplinarity. In *Futures* 36 (4), pp. 515–526. DOI: 10.1016/j.futures.2003.10.007.

Thorne, Frederick C. (1977): The citation index: another case of spurious validity. In *Journal of Clinical Psychology* 33, pp. 1157–1161.

Torre, André (2008): On the Role Played by Temporary Geographical Proximity in Knowledge Transmission. In *Regional Studies* 42 (6), pp. 869–889.

Tsai, Hsu-Hao (2011): Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. In *Scientometrics* 87 (3), pp. 425–450. DOI: 10.1007/s11192-011-0353-6.

Turing, Alan (1950): Computing machinery and intelligence. In *MIND, A Quarterly Review of Psychology and Philosophy* LIX (236), pp. 433–460.

Turoff, M.; Hiltz, S. R. (1982): The electronic journal: A progress report. In *Journal of the American Society for Information Science* 33 (4), pp. 195–202.

Uehara, Masamichi; Takahashi, Ken; Hoshuyama, Tsutomo; Tanaka, Chieko (2003): A proposal for topic-based impact factors and their application to occupational health literature. In *Journal of Occupational Health* 45, pp. 248–253.

Upham, S. Phineas; Small, Henry (2010): Emerging research fronts in science and technology: patterns of new knowledge development. In *Scientometrics* 83 (1), pp. 15–38. DOI: 10.1007/s11192-009-0051-9.

Utterback, James M.; Abernathy, William J. (1975): A dynamic model of process and product innovation. In *Omega* 3 (6), pp. 639–656. DOI: 10.1016/0305-0483(75)90068-7.

Vakkari, Pertti (1999): Task complexity, problem structure and information actions. In *Information Processing & Management* 35 (6), pp. 819–837. DOI: 10.1016/S0306-4573(99)00028-X.

van Dalen, Hendrik P.; Klamer, Arjo (2005): Is Science A Case of Wasteful Competition? In *Kyklos* 58 (3), pp. 395–414.

van den Besselaar, Peter; Heimeriks, Gaston (2006): Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. In *Scientometrics* 68 (3), pp. 377–393. DOI: 10.1007/s11192-006-0118-9.

van den Besselaar, Peter; Leydesdorff, Loet (1996): Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence. In *Journal of the American Society for Information Science and Technology* 47 (6), pp. 415–436.

van Eck, Nees Jan; Waltman, Ludo (2009): How to normalize cooccurrence data? An analysis of some well-known similarity measures. In *Journal of the American Society for Information Science and Technology* 60, pp. 1635–1651.

van Raan, Anthony F. J. (2004): Sleeping Beauties in science. In *Scientometrics* 59 (3), pp. 461–466.

van Rijsbergen, C. J. (1979): Information retrieval. 2nd ed. London, Boston: Butterworths.

Verspagen, Bart (2007): Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. In *Advances in Complex Systems* 10 (1), pp. 93–115.

Vinkler, P. (2007): Eminence of scientists in the light of the h-index and other scientometric indicators. In *Journal of Information Science* 33 (4), pp. 481–491. DOI: 10.1177/0165551506072165.

Wallach, Hanna M. (2008): Structured Topic Models for Language. Thesis. Submitted for the degree of Doctor of Philosophy. Unversity of Cambridge.

Wallin, Johan A. (2005): Bibliometric methods: Pitfalls and possibilities. In *Basic & Clinical Pharmacology & Toxicology* 97 (5), pp. 261–275. DOI: 10.1111/j.1742-7843.2005.pto_139.x.

Waltman, Ludo; van Eck, Nees Jan; van Leeuwen, Thed N.; Visser, Martijn S. (2012): Some modifications to the SNIP journal impact indicator.

Wang, Mingyang; Yu, Guang; Yu, Daren (2011): Mining typical features for highly cited papers. In *Scientometrics* 87 (3), pp. 695–706. DOI: 10.1007/s11192-011-0366-1.

Weingart, Peter (2003): Wissenschaftssoziologie. Bielefeld: Transkript (Einsichten : Themen der Soziologie).

Wendl, Michael C. (2007): H-index: However ranked, citations need context. In *Nature* 449 (7161), p. 403. DOI: 10.1038/449403b.

White, M. D.; Wang, P. (1997): A qualitative study of citing behavior: Contributions, criteria, and metalevel documentation concerns. In *The Library Quarterly* 67 (2), pp. 122–154.

Whittaker, J. (1989): Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. In *Social Studies of Science* 19, pp. 473–496. DOI: 10.1177/030631289019003004.

Wilhite, A. W.; Fong, E. A. (2012): Coercive Citation in Academic Publishing. In *Science* 335 (6068), pp. 542–543. DOI: 10.1126/science.1212540.

Winterhager, Matthias; Schwechheimer, Holger (2002): Schweizerische Präsenz an internationalen Forschungsfronten 1999. Universität Bielefeld. Bielefeld.

Wissmann, Malte; Toutenburg, Helge; Shalabh (2007): Role of Categorical Variables in Multicollinearity in the Linear Regression Model. Department of Statistics, University of Munich (Number 008).

Witten, I. H.; Frank, Eibe (2005): Data Mining. Practical Machine Learning Tools and Techniques, Second Edition. San Diego, Los Angeles: Elsevier Science & Technology Books.

Yan, Erjia; Ding, Ying; Milojevic, Stasa; Sugimoto, Cassidy R. (2012): Topics in dynamic research communities: An exploratory study for the field of information retrieval. In *Journal of Informetrics* 6, pp. 140–153.

Yarime, Masaru; Takeda, Yoshiyuki; Kajikawa, Yuya (2008): Patterns of collaboration in emerging fields of trans-disciplinary science: the case of sustainability science. In *25th DRUID Celebration Conference 2008 on Entrepreneurship and Innovation - Organizations, Institutions, Systems and Regions*.

Yi, Sangyoon; Choi, Jinho (2012): The organization of scientific knowledge: The structural characteristics of keyword networks. In *Scientometrics* 90 (3), pp. 1015–1026. DOI: 10.1007/s11192-011-0560-1.

Zand, D. E. (1972): Trust and managerial problem solving. In *Administrative Science Quarterly* 17, pp. 229–239.

Zhou, Ping; Leydesdorff, Loet (2011): Fractional counting of citations in research evaluation: A cross- and interdisciplinary assessment of the Tsinghua University in Beijing. In *Journal of Informetrics* 5 (3), pp. 360–368.

Zinsser, H. (1940): As I Remember Him: The Biography of R. S. Boston: Little, Brown and Company.

Zitt, Michel; Ramanana-Rahary, S.; Bassecoulard, E. (2003): Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. In *Scientometrics* 56 (2), pp. 259–282.

Zitt, Michel; Ramanana-Rahary, S.; Bassecoulard, E. (2005): Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. In *Scientometrics* 63 (2), pp. 373–401.

Zitt, Michel; Small, Henry (2008): Modifying the journal impact factor by fractional citation weighting: The audience factor. In *Journal of the American Society for Information Science and Technology* 59 (11), pp. 1856–1860. DOI: 10.1002/asi.20880.

Zuccala, A. (2012): Quality and influence in literary work: evaluating the 'educated imagination'. In *Research Evaluation* 21 (3), pp. 229–241. DOI: 10.1093/reseval/rvs017.