Semantic Exploration of Text Documents with Multi-Faceted Metadata Employing Word Embeddings: The Patent Landscaping Use Case

Master's Thesis of

Tatyana Skripnikova

at the Department of Informatics
Institute for Program Structures and Data Organization (IPD)

Reviewer:	Prof. Dr. Ralf H. Reussner
Second reviewer:	Prof. Dr. Harald Sack
Advisor:	Dr. Hidir Aras

01 November 2018 – 31 May 2019

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 31.05.2019

. . . . . .

(Tatyana Skripnikova)

Abstract

The number of written works describing scientific progress is steadily increasing, which necessitates development of supportive tools for their efficient analysis. These documents are characterized not only by their textual content, but also by a number of metadata attributes of various kinds, including any relationships between documents. This complexity makes development of a visualization approach to aid examination of written works a challenging task. Patents exemplify this problem as large amounts of them are studied by companies to gain competitive advantages and guide research and development efforts.

We propose an approach for explorative visualization based on both metadata and semantic embeddings of patent's content. Word embeddings from a pre-trained word2vec model are used to determine similarities between documents. Moreover, hierarchical clustering methods help provide several levels of semantic detail with via extracted relevant key terms. To the best of our knowledge, no existing visualization approach combines semantic embeddings with hierarchical clustering while supporting various interaction types based on metadata attributes.

Our approach makes use of user interaction techniques such as brushing and linking, focus plus context, details on demand and semantic zoom. Because of that, it becomes possible to examine the patterns that result from the interplay between 1) distributions of metadata values and 2) positions in the semantic space.

The visualization concept is shaped by user interviews and evaluated via a think-aloud study with patent experts. During the evaluation we compared our approach to a baseline approach based on Term Frequency - Inverse Document Frequency (TF-IDF) vectors. The usability study indicated that visualization metaphors and interaction techniques were appropriately chosen. Moreover, it showed that the user interface of the prototype played a much larger role in participants' impression than the way patents are situated and clustered. In fact, both approaches resulted in very similar extracted cluster key terms. Nevertheless, the semantic approach resulted in more intuitive relative placement of clusters and better separation of clusters.

Proposed visualization layout, interaction techniques and semantic methods may be extended to other kinds of text documents, i. e. scientific publications. Other embedding methods such as paragraph2vec [61] or BERT [32] could be used to take advantage of contextual dependencies in text above the level of single words.

Zusammenfassung

Die Menge der Veröffentlichungen, die den wissenschaftlichen Fortschritt dokumentieren, wächst kontinuierlich. Dies erfordert die Entwicklung der technologischen Hilfsmittel für eine effiziente Analyse dieser Werke. Solche Dokumente kennzeichnen sich nicht nur durch ihren textuellen Inhalt, sondern auch durch eine Menge von Metadaten-Attributen verschiedenster Art, unter anderem Beziehungen zwischen den Dokumenten. Diese Komplexität macht die Entwicklung eines Visualisierungsansatzes, der eine Untersuchung der schriftlichen Werke unterstützt, zu einer notwendigen und anspruchsvollen Aufgabe. Patente sind beispielhaft für das beschriebene Problem, weil sie in großen Mengen von Firmen untersucht werden, die sich Wettbewerbsvorteile verschaffen oder eigene Forschung und Entwicklung steuern wollen.

Vorgeschlagen wird ein Ansatz für eine explorative Visualisierung, der auf Metadaten und semantischen Embeddings von Patentinhalten basiert ist. Wortembeddings aus einem vortrainierten Word2vec-Modell werden genutzt, um Ähnlichkeiten zwischen Dokumenten zu bestimmen. Darüber hinaus helfen hierarchische Clusteringmethoden dabei, mehrere semantische Detaillierungsgrade durch extrahierte relevante Stichworte anzubieten. Derzeit dürfte der vorliegende Visualisierungsansatz der erste sein, der semantische Embeddings mit einem hierarchischen Clustering verbindet und dabei diverse Interaktionstypen basierend auf Metadaten-Attributen unterstützt.

Der vorgestellte Ansatz nimmt Nutzerinteraktionstechniken wie Brushing and Linking, Focus plus Kontext, Details-on-Demand und Semantic Zoom in Anspruch. Dadurch wird ermöglicht, Zusammenhänge zu entdecken, die aus dem Zusammenspiel von 1) Verteilungen der Metadatenwerten und 2) Positionen im semantischen Raum entstehen.

Das Visualisierungskonzept wurde durch Benutzerinterviews geprägt und durch eine Think-Aloud-Studie mit Patentenexperten evaluiert. Während der Evaluation wurde der vorgestellte Ansatz mit einem Baseline-Ansatz verglichen, der auf TF-IDF-Vektoren basiert. Die Benutzbarkeitsstudie ergab, dass die Visualisierungsmetaphern und die Interaktionstechniken angemessen gewählt wurden. Darüber hinaus zeigte sie, dass die Benutzerschnittstelle eine deutlich größere Rolle bei den Eindrücken der Probanden gespielt hat als die Art und Weise, wie die Patente platziert und geclustert waren. Tatsächlich haben beide Ansätze sehr ähnliche extrahierte Clusterstichworte ergeben. Dennoch wurden bei dem semantischen Ansatz die Cluster intuitiver platziert und deutlicher abgetrennt.

Das vorgeschlagene Visualisierungslayout sowie die Interaktionstechniken und semantischen Methoden können auch auf andere Arten von schriftlichen Werken erweitert werden, z. B. auf wissenschaftliche Publikationen. Andere Embeddingmethoden wie Paragraph2vec [61] oder BERT [32] können zudem verwendet werden, um kontextuelle Abhängigkeiten im Text über die Wortebene hinaus auszunutzen.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background and motivation

Semantic embeddings are used in Natural Language Processing (NLP) to capture relationships between text documents. However, positions and distances in the embedding space are not easily explainable and can hardly be understood by a user by themselves. Additional data dimensions incorporated into the representation of a semantic space provide immense added value. This is especially the case when visually exploring large document collections, where human perception must be aided in the task of finding patterns in data to prevent cognitive overload.

One example of a task in which such exploration takes place is *patent landscaping*. It "constitutes an overview of patenting activity in a field of technology [...] and seeks to present complex information about this activity in a clear and accessible manner" [105]. Patents are an enormously valuable source of technology intelligence. They exemplify the problem at hand because they are text documents with a clearly defined structure, lots of metadata and references to other patent documents. With help of patent landscaping, companies acquire competitive advantages and steer their research and development efforts.

With about 3.1 million patent applications filed worldwide in 2016 [113] and thousands of patent documents subject to analysis for a single domain, an effective approach facilitating the analysis is crucial.

1.2. Objective and research questions

The objective of this work is to provide a solution for the problem of exploration of large document collections. The proposed visualization approach should take particularities of the patent domain into account and therefore be an efficient aid in the task of patent landscaping. At the same time, the proposed approach should be generalizable for application on various kinds of text documents.

The objective presents a number of challenges that have to be addressed. They are described in section 1.3.

Research questions that are being asked in this thesis are:

- How can semantic embeddings be displayed in an transparent and explainable way?

- How can semantic information enhance visual exploration of large document collections?
- How can metadata of various types be combined with the semantic dimension through user interaction?
- Do semantic embeddings provide added value compared with traditional frequency-based representations?

1.3. Challenges

1.3.1. Characteristics of data

1.3.1.1. Vocabulary

Language and especially vocabulary in patent documents deviate significantly from generic written language. Essentially, patents are written in a very abstract way, so that they protect a higher number of potential embodiments (see subsection 3.2.2 for details). This complicates searching for similarities and differences in patent texts.

Whenever the vocabulary used in patents is not too general, it is likely to be very specific. Technical terminology used to describe inventions is very unlikely to be contained in generic text corpora. Moreover, frequency-based text processing methods are susceptible to inaccuracies when a large part of vocabulary consists of rare technical terms. This is why a fine-tuning of the algorithm parameters related to term frequency is necessary. We address this 1) by using a model trained specifically on patent vocabulary (see subsection 5.1.1) and 2) by carefully adjusting parameters for the key term extraction algorithm, especially for cluster key terms as described in subsection 5.1.7.

1.3.1.2. Dimensionality and data types

The documents we are dealing with consist of textual parts and metadata. High-dimensional text representations are necessary to represent content of patents. It is a challenge to map them onto a lower-dimensional visualization space in a beneficial way. Additionally, the visualization approach we aim to develop has to combine semantics gained from text with various visual dimensions derived from metadata of different types. We experiment with multiple dimension reduction techniques as described in subsection 5.1.6.

1.3.1.3. Visual scalability

The datasets that are being analyzed by patent experts can consist of hundreds to thousands of documents. This is not “big data” in the classic sense of the word, but it definitely is on the upper end of the spectrum when it comes to visual representation. The challenge is to develop an approach that works equally well for a wide range of dataset sizes. It

should be able to display thousands of documents in a comprehensive way. For that, the proposed visualization approach has to use screen space wisely and leverage different levels of detail to avoid overwhelming the user. This is why we provide different levels of detail via semantic zooming as described in section 4.1.

1.3.2. Evaluation

The patent domain has been named in [25] as an area where visualization has potential high-impact as a medium for finding causality, forming hypotheses and assessing available evidence. This makes interactive visualization an attractive research topic. At the same time, the nature of the cognitive processes involved makes evaluation difficult.

[23] argues that a great variety of cognitive reasoning tasks exists. Low-level detailed tasks such as compare, contrast, cluster are more clearly defined. High-level complex cognitive tasks include understanding of data trends, uncertainties, causal relationships or learning a domain. No clear definition exists for some of those tasks, so they are challenging to test empirically.

When testing visualization approaches with experts, success in a task may be attributed to an interplay between expert's 1) meta-knowledge, 2) knowledge from other sources and 3) knowledge gained from the presented data. This complicates interpretation of evaluation results further. We address this in our evaluation by having multiple tasks per hypothesis that we evaluate as described in chapter 6.

1.4. Structure of the thesis

After the introduction in the first chapter, in chapter 2 we establish some fundamental concepts our approach builds upon. Among other things, we introduce some definitions from the patent domain that are used throughout this thesis. We then review the state of the art for data visualization approaches that provide means to explore scientific publications or patents.

Then, in chapter 3 we define the framework for the case study to be conducted. The methodology is being established: the frame of reference of the study, the methods for data collection, etc. We justify our choice of semi-structured user interviews for the formative part and think-aloud study followed by a SUS questionnaire for the summative part. We then describe the course of the discussion during interviews with patent experts and summarize our findings and their effect on the development of a visualization concept.

In chapter 4, we first briefly outline the concept for the visualization. We propose a visualization layout consisting of connected views of different kinds of interactive charts: scatter plot, histogram, sunburst with breadcrumbs and detail view. which implement principles of focus + context, brushing and linking, semantic zoom and Shneiderman's information visualization mantra Next, we justify the choice of a two-dimensional visualization space. We then describe how the initial idea appeared and how the concept evolved. Finally, we

briefly cover how we derive semantic representations of documents, cluster them and extract relevant key terms to provide interpretability.

Next, chapter 5 goes into detail about the data processing necessary to prepare patent documents for visualization: preprocessing and cleaning, computing document vectors, dimension reduction, extracting relevant key terms, hierarchical clustering, etc. It also contains a detailed description of the elements in the user interface of the visualization as they were implemented in the prototype. We then elaborate on the interactions between the coordinated views of the prototype.

chapter 6 establishes hypotheses about the visualization approach that pertain to the different visual elements and aspects of interaction. Then, the hypotheses are evaluated through a think-aloud study, which detects usability problems and verifies how suitable the developed approach is for supporting exploration. We confirm that the visualization metaphors and interaction techniques were chosen appropriately. Moreover, the study shows that the user interface of the prototype played a much larger role in participants' impression than the way patents are situated and clustered.

Finally, chapter 7 summarizes the proposed approach and the key findings of this work. We conclude by discussing possible improvements both in a general sense and specifically related to patent domain.

2. Related Work

After having defined the objective and the research questions of this work, in this chapter we first introduce the basic concepts our approach builds upon. We then discuss the state of the art in visualization approaches that deal with large document collections.

2.1. Basic concepts

This thesis unites two domains - information visualization and machine learning, more specifically NLP. This section introduces fundamental concepts from those two research fields that are not new, but serve as a foundation for our approach and a multitude of other works in visualization and machine learning. Moreover, we introduce some terms related to the patent system, which are not a result of some particular existing research as such. They are, however, together with the machine learning and visualization concepts mentioned above, the prerequisites necessary for a solid understanding of this work.

2.1.1. Information visualization

2.1.1.1. Visual information seeking mantra

The visual information seeking mantra by Shneiderman [93] is a seminal concept that has contributed to the success of many powerful visualizations. It consists of three parts:

- overview first
- zoom and filter
- details on demand.

This guideline is crucial for providing optimal bandwidth of the presentation of information .

Overview first means that in the beginning, a more abstract or zoomed out view of a document collection should be presented to the user. The goal at this point is to give a general impression without overwhelming the user. Shneiderman [93] suggests complementing this view with a detail view. Together they yield a focus + context representation (see subsubsection 2.1.1.2).

Zoom allows users to satisfy their interest in some portion of the collection of data. To support this, tools to control the zoom focus (the non-moving point the user zooms

towards) and the zoom factor (the magnitude of the enlargement) are required. Moreover, zooming should be smooth to preserve the sense of position and context.

Filter essentially means applying dynamic queries to the items in the collection and hiding data points that are not in the result set. Filtering helps users to concentrate further on parts of the data they find interesting. Shneiderman [93] argues that updating the display in less than 100 milliseconds is the goal. Such reaction time is necessary to maintain the responsiveness of the system, so that users do not feel uncertain about the result of their action.

Details-on-demand means that the detail view for a single data point should be shown after a request from the user such as clicking on the item. This constitutes the lowest abstraction level of working with a data collection.

2.1.1.2. Focus + context

The basic idea with *focus–plus–context*–visualizations is to enable viewers to see the object of primary interest presented in full detail while at the same time getting an overview–impression of all the *surrounding* information — or *context* — available. Such visualizations are "attention-warped displays". They attempt to use more of the display resource to correspond to the interest of the user's attention[22] - "seeing the trees without missing the forest".

[12] provides an illustrative example of focus + context (see Figure 2.1). They demonstrate a display that has an area with higher resolution nested inside a larger area with lower resolution. A map is presented on the display, with its region of maximal interest (*focus*) inside the area with higher resolution. One can see that only this small region shows street names.

2.1.1.3. Panning and zooming

"*Panning* and *zooming* refer to the actions of a movie camera that can scan sideways across a scene (panning) or move in for a closeup or back away to get a wider view (zooming)" - [45]. This is an ubiquitous interaction form that creates a perception of space and movement in said space. Is in most cases zooming is understood as physical zooming. For comparison with semantic zooming, see subsubsection 2.1.1.4.

2.1.1.4. Semantic zooming

A *physical* zoom changes the size and visible details of objects. A *semantic* zoom, on the other hand, changes the type and meaning of the information displayed by the object by using a semantic transition between detailed and general views of information [73].

A thematically relevant example of semantic zooming is presented by Skupin [98]. He automatically derives the thematic structure of a given domain. As shown in Figure 2.2, he visualized scientific publications from the field of biomedicine.

Figure 2.1.: A focus + context interface. The iconic illustration at the bottom left shows where the focus screen is located. The callout shows the different resolutions of focus and context area. Source: [12]

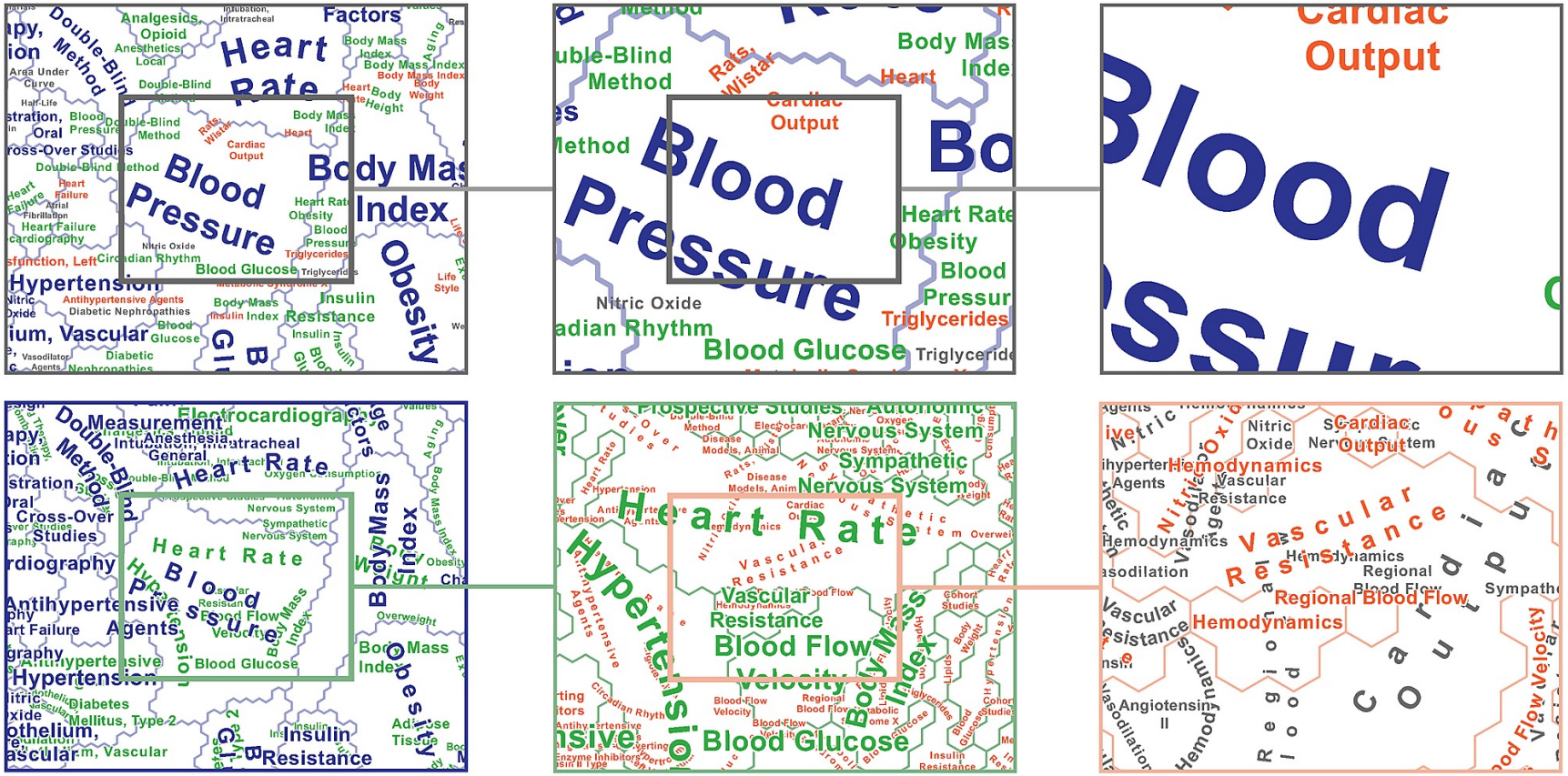


Figure 2.2.: Geometric zooming (top) versus semantic zooming with successive revealing of lower levels of term dominance (bottom). Source: [98]

One can observe that in semantic zooming, the level of detail increases with the zoom level. Specifically, finer subareas with own labels and boundaries become visible. Unfortunately, the zooming itself is not implemented in an interactive way in Skupin's case. Instead, static images are pre-rendered for specific zoom levels. Deciding when to switch between different levels of detail is a separate problem that we address in our work as described in subsection 5.1.7.

2.1.1.5. Brushing and linking

"*Brushing and linking* refers to the connecting of two or more views of the same data, such that a change to the representation in one view affects the representation in the other views as well" - [45]. Multiple views of the same data are usually implemented via different types of visualizations. Common examples include combinations of scatter plots, bar charts, parallel coordinate views and maps. Charts do not necessarily need to be of different types. They may show different dimensions of the data instead (see Figure 2.3).

Brushing means selecting a part of the data in one of multiple views. A selection area is usually formed by dragging the cursor, hence the name *brushing*. Some form of visual indication is necessary to prevent confusion about what exactly has been selected. *Linking* refers to highlighting the selected data points in other view or views. The data is "linked" through the selection. A schematic example of coordinated views with brushing and linking can be seen in Figure 2.4.

2.1.2. Machine learning

2.1.2.1. Word2vec

Word2vec was proposed by Mikolov [72]. It is a neural network architecture for computing continuous vector representations of words in an n-dimensional space. The model learns to predict words based on their context. With enough training data, the hidden layer learns that semantically similar words can be represented with similar vectors. Those word representations are called *word embeddings* and typically have between 100 and 300 dimensions, while less than 50 dimensions usually don't represent the semantics well enough.

Semantic relationships expressed by word embeddings can be of multiple types. [72] evaluated semantic and syntactic relationships in question-answer pairs such as city-in-state (i. e. Chicago - Illinois), man-woman (i.e. brother - sister), opposite (ethical - unethical), nationality-adjective (Switzerland - Swiss). If two words are close in the embedding space, it might also mean that they are synonyms or often appear together in texts. An illustrative overview of concepts behind word2vec, its training and parameters can be found in [5].

Figure 2.3.: Correlation matrix of a well-known Iris dataset as an example of brushing and linking. Linked views are all of the same type, which in this case is a scatter plot. Brushed view is in the upper-left corner. Image source and demo: [18]

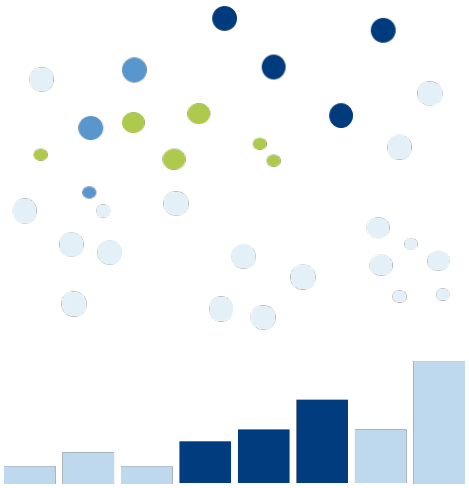


Figure 2.4.: A histogram and a scatter plot coordinated through brushing and linking. The data points that do not belong to the current selection are grayed out in both views.

2.1.2.2. t-SNE

t-SNE is a dimension reduction technique proposed by [60]. The aim of the technique is to map high-dimensional data to a two- or three-dimensional space. At the same time, the structure of the high-dimensional data should be preserved as much as possible. In other words, t-SNE defines a probability distribution in high-dimensional space and tries to preserve it for low-dimensional space using *gradient descent*. Gradient descent is an iterative optimization algorithm which aims to find a minimum of a function by taking small steps in the direction of the steepest decline. The optimization is initialized randomly and is in the case of t-SNE defined by a non-convex cost function, which means that risk of getting stuck in local minima exists. It is, however, completely acceptable to run the algorithm multiple times and choose the best result. t-SNE succeeds in preserving both global and local structure of the data, which makes clusters visible at several scales. An excellent interactive overview of t-SNE's parameters (especially perplexity) and their influence on resulting behavior of the algorithm can be found in [109].

2.1.3. Definitions from the patent domain

This section contains some basic knowledge about the patent system which is a prerequisite for an understanding of the patent visualization tools related to this work (described in section 2.2), and also for an understanding of our own approach (described in chapter 4). We show how patent documents are structured. We also define some terminology from the patent domain that we use throughout this work: patent family, citations and IPC classes.

2.1.3.1. Structure of a patent document

Each patent document possesses the following attributes:

- *Application number* - a unique identifier, starts with country code of the registration country. Example: US-5448677-A.
- *Country code* - stands for the code of the patent office the application was submitted to. While some countries, like the US, have their own designated patent offices, there are patent offices, such as the European Patent Office, that allow patents to be valid in multiple countries. In those cases, terrestrial validity of patents is a complex topic. For the purposes of simplification, we only consider the code itself (e.g. US or EP) in this work.
- *Priority date* - the date when the priority patent was submitted (see subsubsection 2.1.3.2 for details on priority). For our purposes this date is considered as the creation date of the patent.
- *Assignees* - a list of one or many individuals (inventors) or institutions to which the rights to the invention belong to. It is a categorical attribute.
- *IPC classes* - a list of one or many codes from the IPC classification (see subsubsection 2.1.3.4 for more details) describing the thematic areas of technology the described invention is related to. It is a categorical attribute of a hierarchical nature.
- *Citations* - a list of patents (identifiable by their application numbers) cited by the given document.
- *Family identificator* - a number that all members of one patent family share.
- *Title* - a text attribute describing the invention very briefly . It often does not contain enough useful information if used by itself, but adds some clarity when combined with the abstract.
- *Abstract* - by analogy with scientific publications, it is a brief summary of the invention.
- *Claims* - detailed description of the invention and its aspects that should be protected (claimed) by a patent. This is the field with the most textual information as it can be hundreds to thousands of words long. Our main data source, Google Patents Public Datasets [111], only provides claims for patents registered in the US.

This list is not extensive and only includes fields that are relevant to this work. For the above-described fields, we distinguish between *textual content* of a document (title, abstract and claims) and the *metadata* which includes all remaining information.

2.1.3.2. Patent family and priority document

Patents can be assigned to the same *patent family*. Protection of intellectual property for a patent is restricted to the country the application was submitted to. Therefore, many inventions are registered in patent offices in multiple countries. The patents covering one invention across several countries constitute a family. The earliest patent from a family is called a *priority document*. Families can also be registered in the same country when they describe different aspects of the same invention.

2.1.3.3. Forward and backward citations

A patent may include a list of tens to hundreds of citations defined by their application numbers. When a new patent application cites an already existing patent, it indicates that the cited patent is already known to the applicant [31]. The older patent in this case is considered *prior art*. The new application must provide claims that are novel and non-obvious in the view of the prior art. It is in the interest of the applicant to show (through a citation) that they have thoroughly studied already existing patents. This is analogous to scientific publications where the authors have to review state-of-the-art before proposing novel approaches.

The citations directly defined in a patent, i. e. links to older documents that are being *cited*, are called *forward citations*. *Backward citations* are the same links from the point of view of the older patent. They show the patents *citing* the current one. As it is impossible to know how a patent will be cited in the future, an explicit list of backward citations does not exist and has to be assembled by reversing forward citations.

2.1.3.4. International Patent Classification

Each patent is assigned at least one, but usually multiple IPC codes. The IPC hierarchy breaks the whole of humanity's patented technological knowledge down into thematic areas the inventions pertain to. One IPC class is an alphanumerical code which is hierarchical in nature and is based on prefixes, i. e. it starts with one letter and with addition of further symbols corresponding new nodes in the hierarchy tree appear. The tree structure is of the constant depth of five levels and, as we learned in the expert interviews (described in subsection 3.2.5), those levels have own names and are composed according to certain rules (see Table 2.1).

Name	Format	Example	Example title
Section	One letter	H	Electricity
Class	Two-digit number	H04	Electric communication technique
Subclass	One letter	H04N	Pictorial communication, e.g. television
Group	One-to-three-digit number	H04N5	Details of television systems
Subgroup	Two- or three-digit number	H04N5/76	Television signal recording

Table 2.1.: Structure of IPC classes

2.2. State-of-the-art visualization approaches

After all prerequisite concepts have been introduces, we review the state of the art with regard to explorative visualization approaches.

Federico et al. [38] surveyed 21 existing visualization approaches for patents and 109 for scientific documents such as papers, focusing on non-commercially available tools.

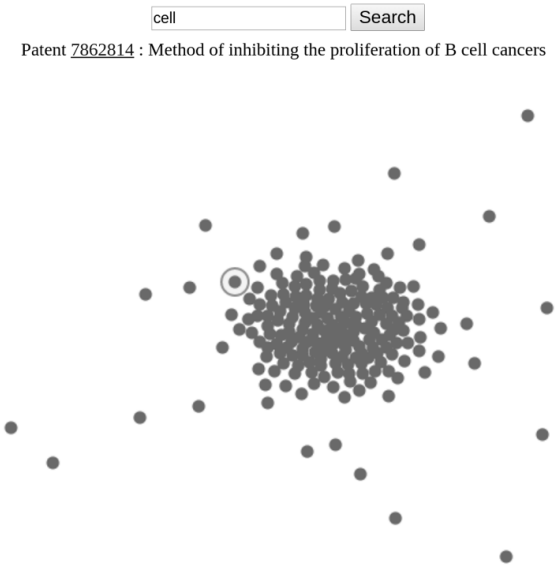


Figure 2.5.: The visualization proposed by [52]. Output of a query for "cell". The point in a circle shows the currently selected patent.

They distinguish between four data types that can be visualized: *text*, *citations*, *authors* and *metadata*. We focus our review of state-of-the-art on approaches that visualize 1) text alone, 2) other data types alone and 3) text in combination with other data types. A separate section is devoted to a group of comparable themescape-based approaches. In the following, a selection of works discussed by Federico et al. is expanded by some other approaches which they did not include.

2.2.1. Text-based visualizations

Johnson et al. [52] present a similar visualization approach to ours with regard to processing the text data for the visualization. They use a word2vec model trained on ca. 1.5 million patent texts and compose document embeddings through averaging of word embeddings as well. They, however, only focus on textual content and disregard visual representation of metadata. An overview of the whole dataset is not provided. Instead, one needs to query the data by keywords and only subset corresponding to the query is then shown (see Figure 2.5. The only interaction available apart from querying is the selection of a single patent by clicking on the corresponding point, so that patent details are displayed. The approach proposes no method to automatically label data points to provide an overview.

Skupin [97] [96] applies a cartographic approach to create maps of non-geographic information, more specifically, conference abstracts. In a successor work (see Figure 2.6(a)), Skupin et al. [98] visualize medical publications based on MeSH terms, which are analogous to tags assigned to scientific articles. In all of those works, a type of Artificial Neural Network (ANN) called Self-Organizing Map (SOM) is used. The network is trained on a term-document-/MeSH-document matrix. This way, each neuron is assigned multiple terms as labels. In [98], the neurons are then clustered in the following way: "if two

neurons are neighbors in the two-dimensional neuron lattice and they share the same top-ranked label term, then their boundary is dissolved, thus forming a larger polygon, a neuron label cluster" [98]. The same procedure is duplicated for the second and third top-ranked label terms, and the resulting three clusterings are then overlaid on the same map and distinguished by color. Cluster size declines with the dominance of the term, which means the end result displays multiple levels of semantic detail.

In [97] and [96], Skupin uses an alternative clustering method - hierarchical clustering. From the tree structure built on neuron similarities he derives three to five clustering levels as shown in Figure 2.6(b). This approach served as an inspiration for our own hierarchical clustering based on similarities between patent documents.

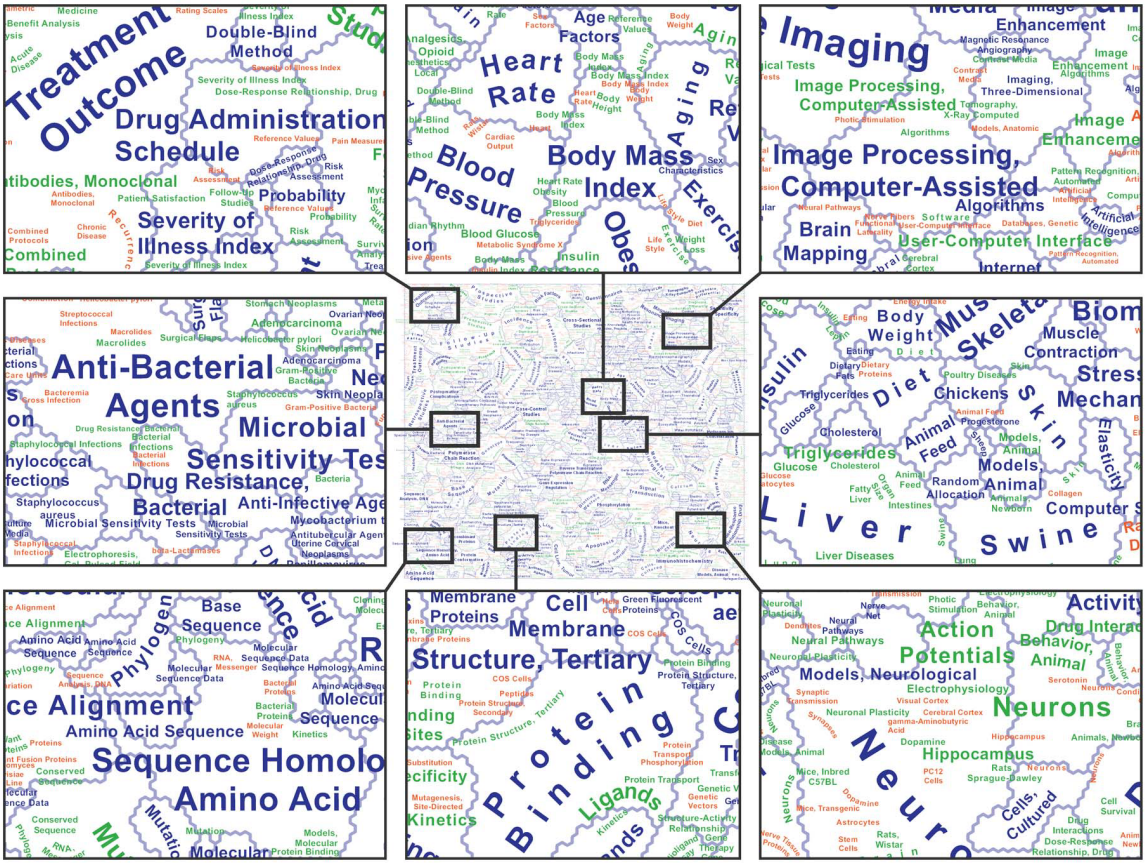
Our approach builds upon Skupin's ideas and aims for a comparable result, while the features of the data and the processing methods are different. Visualizations proposed by him are either static or provide a basic zooming interaction while limiting the richness of display. We, on the contrary, specifically focus on supplementing a hierarchical themescape with interactivity for effective exploration.

Choo et al. [27] present UTOPIAN (**U**ser-driven **T**opic modeling based on **I**ntera**c**tive **N**onnegative Matrix Factorization). They perform topic modeling based on a bag-of-words representation of a document. A hard-clustering algorithm is then applied to the documents, which means that each document is assigned to only one topic. A modified version of t-SNE is utilized to draw a node-link diagram with topics/clusters distinctly separated as shown in Figure 2.7. Edges are drawn between pairs of data points whose distances are below a user-specified threshold. Most importantly, the approach gives the user a high level of control over the topic modeling result. Merging or splitting topics, creation of a new topic based on a specified document or a certain keyword are possible. Choo's approach influenced our initial idea of drawing a fully connected graph of documents which would dynamically reform itself after some documents are filtered out (see subsection 4.2.3 for details). We, however, did not pursue this idea further because of performance considerations.

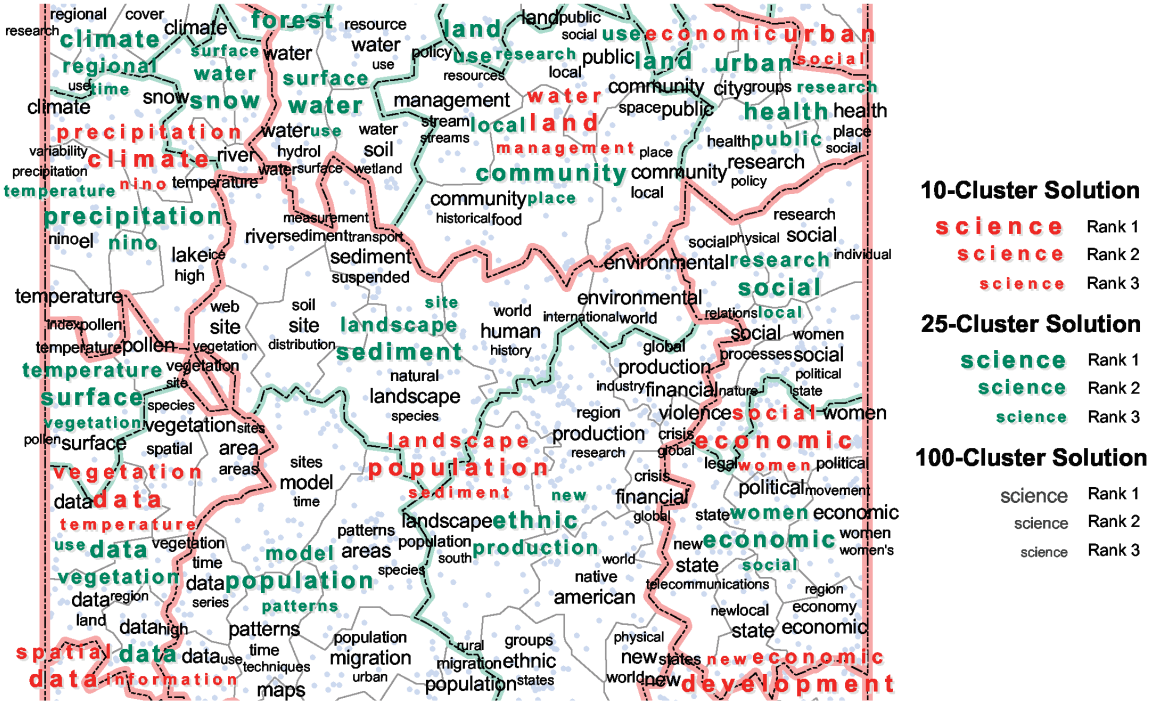
2.2.2. Visualizations based purely on metadata

Patent data is distinguished by a significant amount of metadata attached to each record: hierarchical classification, assignees, citations, patent family information, etc. Many works deal with one or a number of these attributes [115] [112] [46] [40] [39] [24] [2]. However, Federico et al. emphasize that "only few works adopt, refine, or develop techniques for visualizing classification data. Other data types are just ignored in most approaches" [38]. Because of this, we aim to derive value specifically from the hierarchical representation of IPC classification data.

Wittenburg et al. [112] make extensive use of metadata for faceted visualization with what they call embedded bar charts. They order the company, decade of filing date, country and IPC class vertically over each other and represent the distribution of values within those attributes through widths of blocks as seen on Figure 2.8. The embedded bar charts use negative space within the blocks to display temporal development of a company's



(a) Zoomed-out view of a complete map of medical literature and detailed views of some regions. Blue, green and red labels indicate clusters derived from the first, second and third top-ranked label term, respectively. Source: [98]



(b) Visualization of conference abstracts with simultaneous overlay of three levels of a hierarchical clustering. Source: [97]

Figure 2.6.: Skupin's maps of knowledge domains

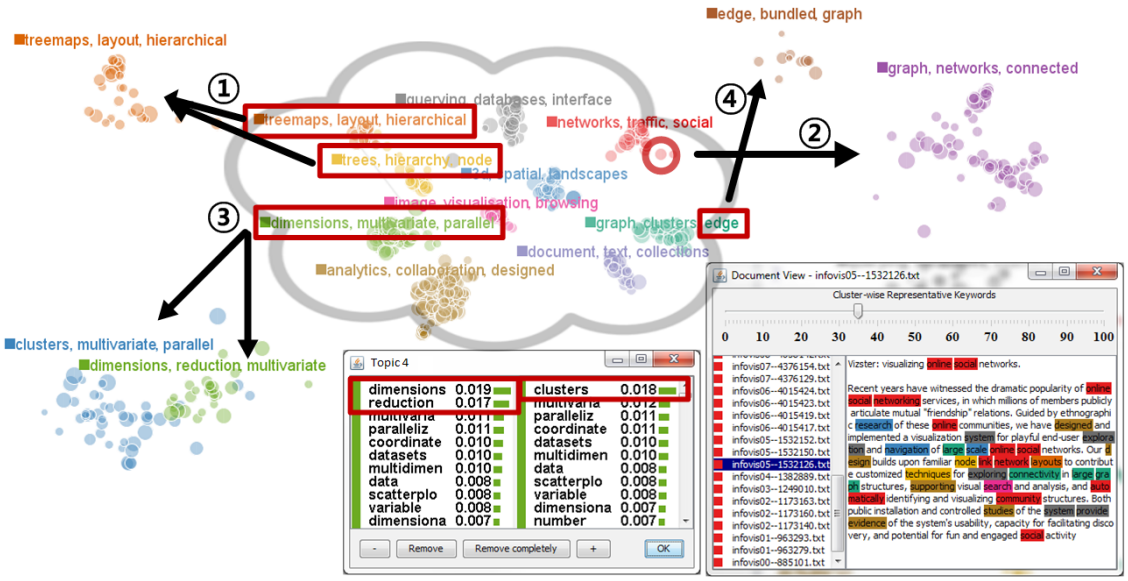


Figure 2.7.: UTOPIAN by [27]. Given a scatter plot visualization generated by a modified t-SNE, it provides capabilities for 1) topic merging, 2) document-induced topic creation, 3) topic splitting and 4) keyword-induced topic creation. The user can adjust topic keyword weights (bottom-middle) and see representative keywords in the document viewer (bottom-right).

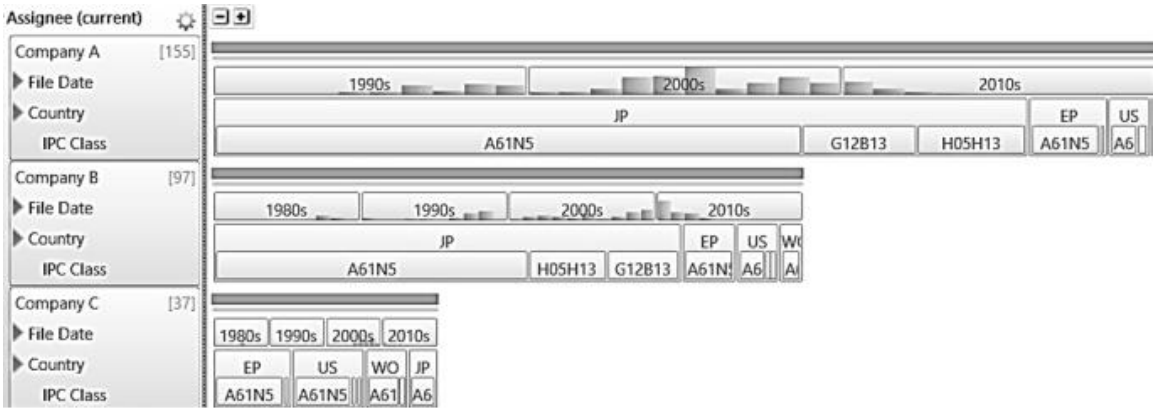


Figure 2.8.: A visualization layout proposed by [112], a so called embedded bar chart. The distribution of metadata attributes in the dataset is represented by a hierarchy of attributes: assignee, then date of filing, then country, then IPC class.

patenting behavior. Unfortunately, Wittenburg's approach results in a cluttered view and therefore lacks visual scalability when many companies and especially IPC classes are present in the data. Nevertheless, we build upon their idea about displaying distributions of metadata attributes in a stacked order.

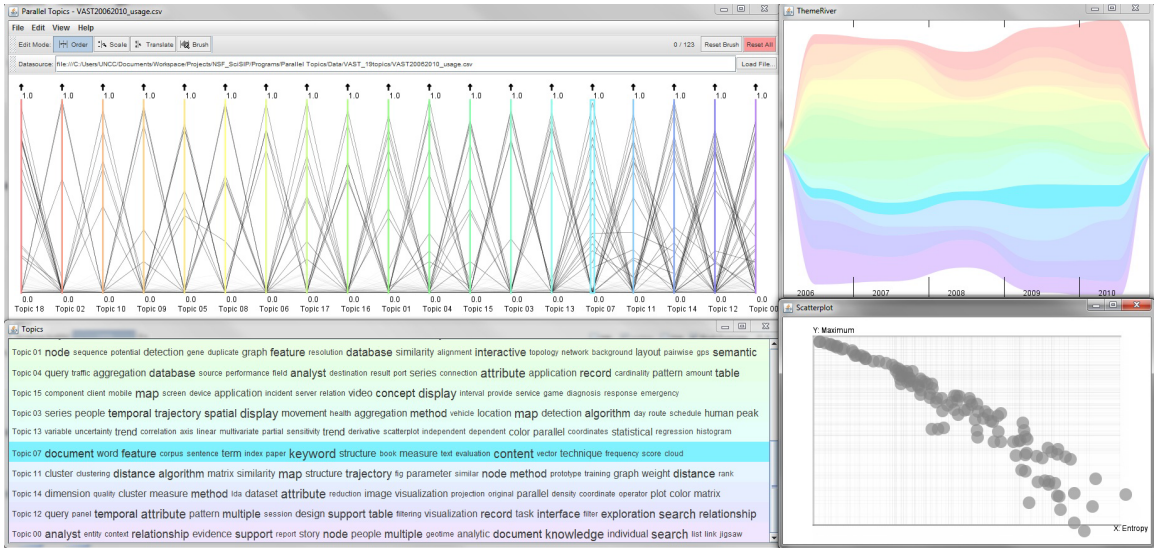


Figure 2.9.: ParallelTopics by [34]. Top left: Document Distribution view, top right: Temporal view, bottom left: Topic Cloud, bottom right: Document Scatterplot.

2.2.3. Visualizations based on text in combination with other data types

Many approaches capture thematic similarities between documents with help of topic modeling [75] [42] [34] [51]. They position a document depending on the degree to which it belongs to the corresponding topic. For example, Dou et al. [34] uses the parallel coordinate metaphor to present a probabilistic distribution of a document across pre-detected topics as seen in Figure 2.9. With an interactive ThemeRiver view they present the temporal development of the topics. Lastly, they use a scatter plot to show the distribution of single-topic vs. multi-topic documents. Pie glyphs within the scatter plot describe the topical contribution to a specific document. We consider the parallel coordinate plot a suboptimal choice to represent a large number of similar dimensions such as >10 topics. Even though the topic axes are ordered by similarity, the positive and negative correlations between topics become spread over the >10 vertical axes, which makes them hardly perceptible. Nevertheless, this approach served as an inspiration for our own glyphs in the form of a pie chart.

Jiang et al. [51] present a comparable approach to Dou et al. as seen in [**fig:jiang**]. They detect tens of topics using a hierarchical topic model. Each topic is then represented as a vocabulary-length feature vector where each dimension corresponds to the word's probability in a latent semantic space. The topic vectors are then reduced to two dimensions via Multidimensional Scaling (MDS) and are shown in the form of a scatter plot. This approach comes close to our idea of making patterns in the semantic space visible. Unfortunately, it does not provide a representation of how separate documents relate to topics. It also involves only the temporal dimension of documents and no as an addition to their textual content, no other metadata is represented.

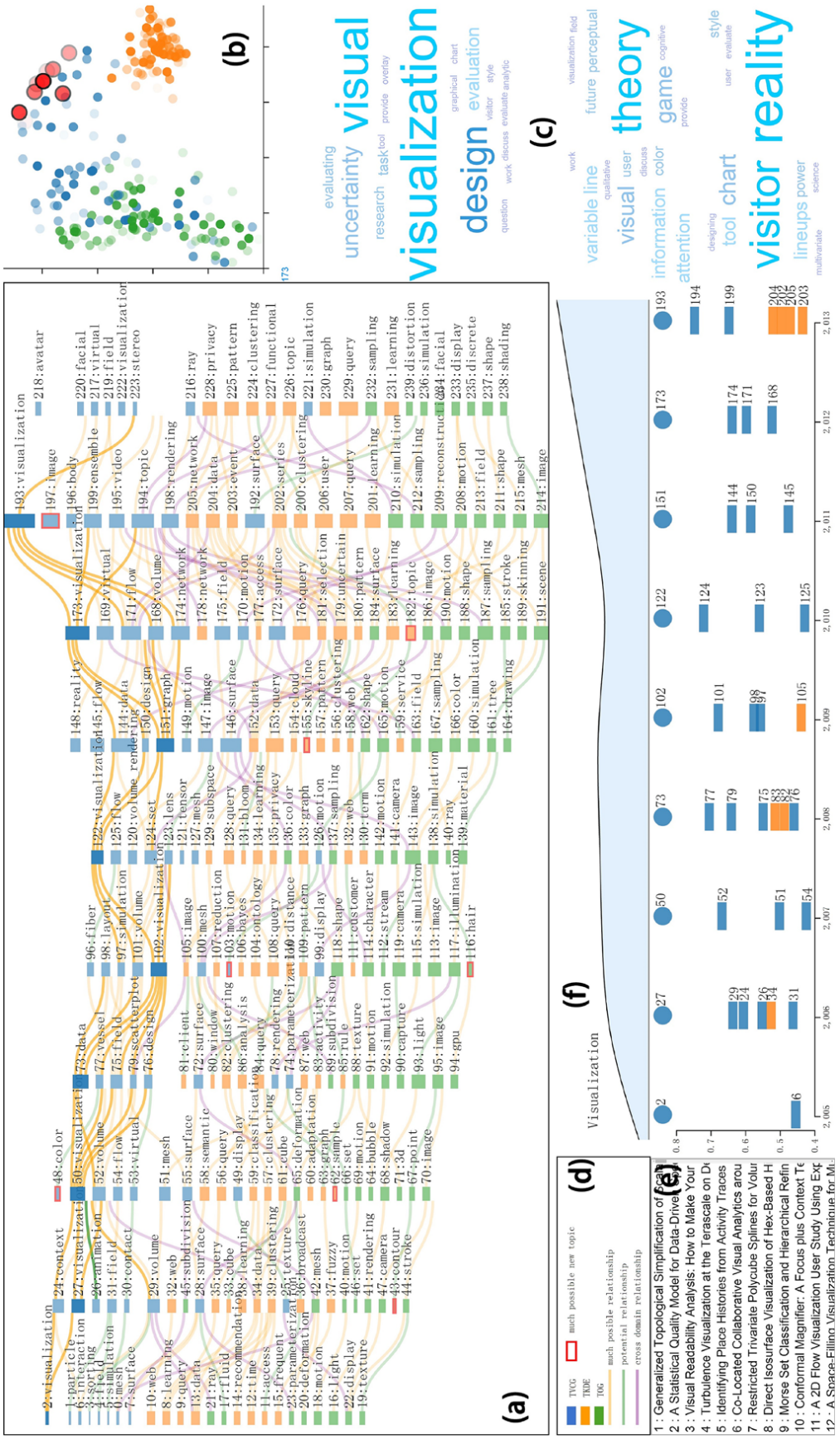


Figure 2.10.: A visualization approach by [51]. a) sankey diagram presenting the temporal development of numbered topics; b) scatter plot showing the topics in 2D space; c) word cloud of the selected topic and subtopics; d) legend; e) titles of papers belonging to the selected topics; f) stream diagram illustrating the topic trend with a scatter plot to represent topic similarities.

2.2.4. Themescapes

Visualization approaches that produce themescapes constitute a separate noteworthy category. They include commercial tools such as VxInsight [19], Thomson Reuter's Aureka and STN's AnaVist. Information about those tools is limited because of their cost, but [91] provides an extensive comparison. There are also some non-commercial approaches such as IN-SPIRE [47]. All of those approaches utilize the metaphor of points in a landscape comprised of "mountains" and "valleys" as seen in Figure 2.11. Mountains group patents with similar textual content via word-frequency-based similarity metrics. The height of a mountain peak corresponds to the document density in the area. A peak is usually labeled with a list of automatically extracted relevant terms. Skupin's work takes advantage of the map metaphor as well, but does not completely fit into this category because he does not use the third dimension to represent the amount of documents in a cluster.

In most themescape-based approaches, the user can highlight points on the landscape which correspond to a certain author, patent assignee, time period, country, etc. Thus, a distribution of metadata values can be explored. Moreover, coordinated views supplement the main landscape view by providing statistics in form of histograms, co-occurrence matrices, pie charts, citation graphs, etc. (see Figure 2.12 for an example from AnaVist).

The little information that is publicly available about the commercial themescape-based tools seems to indicate that all of them use document representations based on frequency and/or distribution of words. In such classical machine learning methods, words are treated like indices in a dictionary and there is no concept of context or similarity between words. In our work we compare one such approach (TF-IDF document vectors) with a newer neural-net-based approach that takes semantics of words into account.

2.2.5. Summary

Many approaches use topic modeling as a way to give meaning to the positions in the visualization space. Others use similarity metrics based on word frequencies to cluster documents. Only one approach (Federico et al. [38]) uses semantic word embeddings instead of word-frequency-based features.

Very few works handle patent classification data, which is why we explicitly focus on finding an appropriate visual metaphor for IPC classes.

Ultimately, we are unaware of any approach that 1) relies on semantic embeddings to show local and global structures within a dataset, 2) organizes themes into a conceptual hierarchy via clustering and at the same time 3) enables exploration of document metadata through additional visual dimensions or interaction techniques. This is the research gap we address in this work.

Figure 2.11.: A patent landscape map about graphene produced with Aureka. Highlighted are Samsung's patents published in 2013 and 2014. Image source: [41]

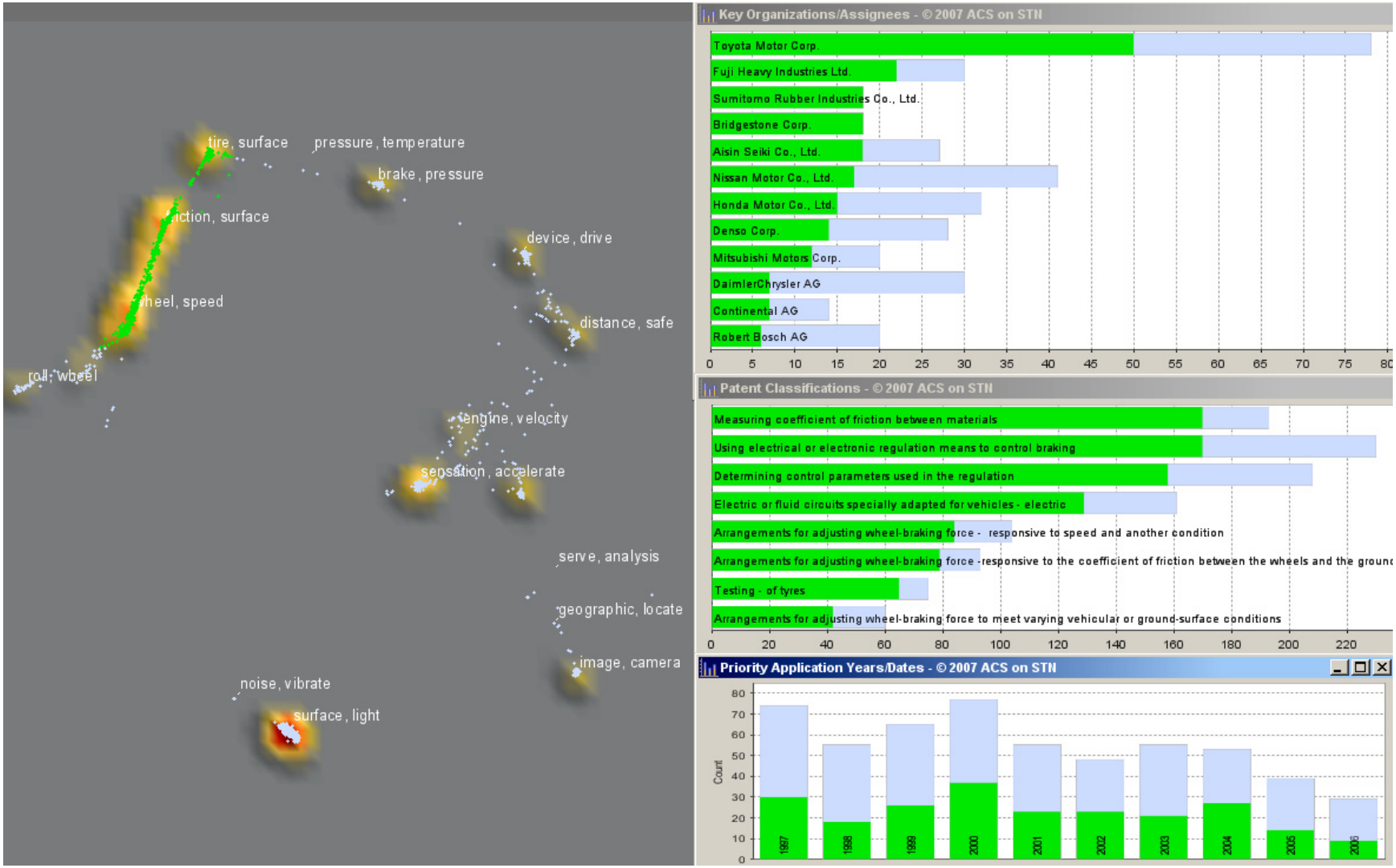


Figure 2.12.: Screenshot of STN AnaVist, a commercial tool for patent landscaping. A selected subset of the data is highlighted in green. Image source: [91]

3. Case study

After we have covered the basic concepts and reviewed the existing visualization approaches, we define the framework for the case study which directs our solution and, later, helps to evaluate it.

3.1. Design of case study

A *case study* is an empirical method aiming at investigating contemporary phenomena in their context. While it does not uncover causal relationships as well as a controlled experiment would, it provides deeper understanding of the studied phenomenon and is flexible [90]. Nevertheless, a case study has to be carefully planned.

3.1.1. Plan for the case study

According to [87], a plan for a case study has to include:

- Objective—what to achieve?
- The case—what is studied?
- Theory—what is the frame of reference?
- Research questions—what to know?
- Methods—how to collect data?
- Selection strategy—where to seek data?

In this section we answer those questions.

Tools already exist to support exploration of corpora of text documents in general and the patent landscaping process specifically. Therefore, the case study has an *improving* objective.

The objective and the research questions of this work have already been covered in detail in section 1.2.

The *case* that is being studied is the task of patent landscaping.

The study operates under the assumption that semantic embedding results in a similarity measure that is meaningful for human perception. This assumption defines the *frame of reference*.

Next, methods for data collection and data selection strategy have to be defined. We use *first degree* data collection techniques, more specifically, user interviews and a think-aloud study complemented by a SUS questionnaire. The participants for those studies are experts from the patent domain.

First degree data collection techniques are methods in which the researcher is in direct contact with the subjects and collects data in real time[62]. Second and third degree techniques mean collecting data without direct participant interaction or using data that already exists, respectively. A comparison of techniques of various degrees of access can be seen in Figure 3.1.

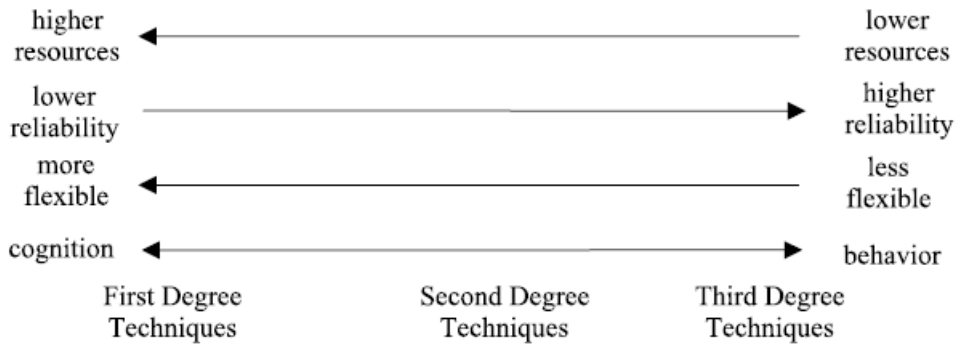


Figure 3.1.: Cost, reliability, flexibility and cognition vs. behavior compared. Source: [62]

First degree techniques require more time and effort from both researcher and study participants. This is due to the fact that they tend to produce a large amount of data that needs processing. On the positive side of this trade-off, first degree techniques provide the researcher with more flexibility and control over data collection. Most importantly, first degree techniques allow the researcher not only to understand *how* the task is performed (behavior), but *why* (cognition). The downside is that the gathered data relies on imperfect human recollection, so care must be taken if complete accuracy of reported facts is important. Since we are interested in overall cognitive processes instead of minutia, we implement first degree techniques in our case study.

3.1.2. Procedures for data collection

After having settled on using first degree data collection techniques, in this section we define how the data is to be collected. This includes a detailed discussion of the user studies that produce the data to collect.

Formative study

First, a formative study in the form of expert interviews serves to understand scenarios in the patent landscaping task and formulate the requirements. The purpose of this first stage of data collection is to acquire subjective, qualitative results since actual human experiences provide valuable insight into benefits and drawbacks of existing solutions.

Accordingly, *semi-structured* interviews are chosen to understand the users' mental model of the task. A semi-structured interview consists of a mix of open and closed questions. This type of interview is common in case studies [90]. It allows the researcher to follow the natural development of the conversation, improvise and explore the subject at hand while making sure that all relevant topics are addressed. We discuss the results of the formative study in section 3.2.

Ideation

After the first data collection stage, a concept for the visualization itself must be developed. The concept is influenced by the insights gained in the formative study. The result of the decisions made at this stage is a digital mockup representing the future interactive prototype (see chapter 4 for details). An implementation phase that follows consists of data preprocessing, applying chosen semantic methods and implementing chosen interaction techniques in a usable interactive prototype (see chapter 5 for details). After the first proof-of-concept prototype is complete, a short feedback meeting with the potential users takes place . This helps gather first reactions to the concept and, when necessary, adjust further iterations.

Summative study

An evaluation of the approach is concluded by the second data collection stage. The execution of this stage is covered . One of the purposes of the second data collection stage is to uncover cognitive problems and mismatches between the user's mental model of the task and the proposed system. The second purpose is to evaluate the impact of semantic embeddings as compared to a traditional approach such as TF-IDF features as document vectors. To do that, participants are divided into two groups. One group evaluates the traditional approach first and the approach with semantic embeddings afterwards. For the second group, the order is reversed.

To gather direct qualitative feedback about usability, a think-aloud study is planned. The think-aloud approach has its roots in cognitive psychology and is scientifically established [95] [36]. It was originally applied in studying short-term memory processes. Two embodiments of the think-aloud method exist. Ericsson et al. [35] keeps the influence of the experimenter on the outcome to a minimum with rigid procedures. Contrarily, [16] et al. approach the experiment as a dialogue. The participant is still encouraged to talk most of the time. The researcher mostly listens and acknowledges what is being said, but is allowed to ask questions or intervene in case the participant is lost or a bug in the tested system prevents further progress. The two techniques were evaluated in [57]. The outcome shows that the subjects' evaluations were consistent between methods. However, the subjects completed more tasks and felt less lost with approach from Boren et al. We therefore choose this embodiment of think-aloud for the qualitative part of our summative study.

For quantitative feedback, we measure the usability score resulting from the use of the prototype. This helps verify that the visualization based on the proposed approach is easy to use and satisfies the requirements. For that, SUS questionnaire [21] is chosen. It

is preferred to other questionnaires such as Computer System Usability Questionnaire (CSUQ) [65] and Questionnaire for User Interface Satisfaction (QUIS) [26] because it produces reliable results even with small number of participants [107]. Moreover, it is short, simple and addresses different aspects of user's reaction to the tested system as a whole instead of its specific features. Alternating positive and negative questions ("I thought the system was easy to use" vs. "I found the system unnecessarily complex") require attention from participants and provide more robust results. [59] proposes to combine a SUS scale with a follow-up question about reasons for the given rating to derive further qualitative insights. We follow this suggestion.

Study subjects

Experts with experience in patent landscaping from FIZ Karlsruhe serve as subjects of the study. According to Nielsen [78], about 70% of the insight can be learned from three participants. Additional participants, especially those after the fifth one, bring merely diminishing returns. For this reason, the study is kept small with 3 participants for user interviews and 4 for the think-aloud and SUS part.

Analysis of the data collected during the second data collection stage allows drawing conclusions regarding the objectives of the study. Those are described in section 6.3.

3.2. Interviews

Since the formative study defined in the previous chapter had a direct impact on the development of our approach described in chapter 4, we cover the study itself and its results in this section. First, we describe the organizational and methodological aspects of conducting a semi-structured interview, which is an integral part of the formative study. After the organizational and methodological aspects follow the descriptions of conversations with patent experts. We conclude by a summary of how the interviews shaped our understanding of the patent domain and, subsequently, how they influenced the development of a concept for our approach.

3.2.1. Procedure

Three patent experts were interviewed in a semi-structured format as defined in [87]. After greeting the participant, each interview started with some warm-up questions about participant's background. Afterwards followed the main part with the most important questions about the domain of patent landscaping itself. The interview was concluded by cool-down phase with more general questions. The questions covered relevant aspects of a patent landscaping process such as data quality, usage scenarios and working with different abstraction levels. A full questionnaire for the interview can be found in section A.1.

Best practices for conducting user interviews were studied and adopted to the best of interviewer's ability. Here we list the guidelines that we followed based on [82]:

- Before scheduling the interview, ask participants if you are allowed to record them talking. It is virtually impossible to actively listen and steer the conversation while taking extensive notes. An audio recording is usually sufficient. Don't forget to check recording equipment before first interview starts.
- The questions should not assume a certain point of view. For example, "How do you feel about X?" is better than "Why is X bad?". Ask even if you think you know the answer, you might be surprised.
- Show that you understood what the interviewee is saying and ask for clarification. For example, "You said X, could you please tell me more about it?".
- If questions come up while the interviewee is speaking, don't interrupt them. Write the question down and follow up later.
- Speak slowly, don't show hurry. If the time is running out, prioritize.
- Pay attention to interviewee's body language and try to imitate it when appropriate. Try to prevent defensive poses such as crossed arms.
- Leave long pauses after the interviewee's replies. Silence is mildly uncomfortable and serves as a prompt to keep talking. It also gives the participant a chance for contemplation, allowing them to formulate additions to their last thought.
- After the main part of the interview is finished, ask the participant if they have questions for you or would like to tell you something you both had not discussed yet.
- After completing the interview thank the participant, stop the recording and note the main topics of the conversation.

All participants were to some degree familiar with STN AnaVist, which is an interactive visualization software specifically created for use in the patent domain. As seen in Figure 2.12, it displays a patent map with labeled clusters and allows the users to select an area to compute statistics about the patents in that area as compared to the whole dataset. Prior exposure to STN AnaVist most probably shaped experts' expectations for a patent landscaping tool.

The interviews provided valuable insights into workflows and mental processes of patent experts. In the the descriptions of the conversations we only elaborate on discussion points that are relevant to the development of our approach. The issues not covered here nevertheless significantly contributed to our understanding of the patent domain.

We refer to the participants by the letters of the Greek alphabet and singular "they" for both genders to preserve their anonymity. The grammatical form of singular "they" also applies for any (potential) users we mention throughout this work.

3.2.2. Participant Alpha

Participant Alpha has the most experience with patent landscaping. They composed patent landscape reports and are very familiar with the landscaping tool STN AnaVist and presented it to clients.

Participant Alpha characterized creating a patent landscape as an iterative process. It consists of multiple feedback loops that run until converging to a satisfactory result.

The first feedback loop involves understanding the needs of the client better. It starts when a client commissions a patent analysis and explains their requirements to the expert. The mutual understanding of the task is difficult to achieve, especially when it evolves based on illustrative results. Therefore, after the patent expert presents the client with the result of the current iteration, the client may influence the focus of the analysis. The expert incorporates the client's suggestions into the further workflow.

The second feedback loop concerns the level of abstraction in the query to the patent database. Participant Alpha pointed out that the patent attorneys often use very generic vocabulary compared to scientific publications such as papers. This allows the claimed invention to be protected in a wider variety of embodiments. Patent offices work against this tactic by demanding a sufficient level of detail to prevent claims from being too broad. As a result, patent expert sometimes has to experiment with making terms of the search query more or less generic. If a query consists of parts A, B and C, a combination of generic search terms for A and B and specific terms for C might be followed by a combination of specific terms for A and C with generic terms for B. Thus, possible combinations of generic and specific terms for parts of the query are tested iteratively until the query result is satisfactory. The optimal level of detail for each part of the query constitutes a substrategy, and such substrategies are merged in the final query.

Participant Alpha highlighted the importance of uniform names for assignees and inventors. They reported a recent "information flood" from Asia, especially from China, which necessitates uniform rules for transliterating proper nouns. According to the participant, uniform names of patent assignees are a distinguishing feature of high-quality datasets that is required by clients. Non-uniform names make aggregating data, i. e. counting, unreliable.

The participant made a distinction between two kinds of patent landscaping. One involves looking at a patent set in a quantitative way through the set of metadata attributes. It helps answer questions such as who are the biggest competitors, since when they have been active, how many applications do they have, in which countries are they active. The second kind involves better understanding of the technology domain and ability to subdivide it into subdomains. The participant named freedom-to-operate research as a possible scenario for this kind of analysis. In this area they saw potential for use of semantic methods such as the contribution of this thesis.

The participant described multiple definitions of a patent family. The most widely accepted one and also the broadest one is "simple family", which constitutes a group of patents associated with the same priority document. Simple family may contain patents with claims that differ significantly because they protect different parts of the same invention or take into account regional differences. Creators of patent databases such as Derwent also create their own definitions of family which are more narrow and typically contain almost identical patents only.

Aside from the freedom-to-operate scenario, the participant also mentioned 1) whitespot analysis, 2) searching for cooperation partners and 3)"licensing out" patents that are

not in company's core portfolio as valid scenarios where a patent landscaping tool has successfully been used or might be useful. If one searches for a widespread technology, whitespots can be recognized where a few data points are on their own and not in any big cluster.

Participant Alpha emphasized strongly the importance of data quality. They reported that a patent in a full-text database might be about 100 pages long, while same patent in an added-value database, such as Derwent, might be about 2 pages. This difference is explained by the fact that texts in an added-value database are rewritten by professional writers to a more concise and understandable form. Added-value databases are used as a default option during a patent search. Full-text databases are only searched when necessary.

While on the topic of patent classification systems, participant Alpha reported that IPC is used by virtually all patent offices worldwide and covers about 98% patent documents. Though Cooperative Patent Classification (CPC) has more meaningful and better structured hierarchy, it is assigned to only 40% of documents and is therefore unreliable if used by itself. The participant named subclass level (i. e. A61K), main group (i. e. A61K6) and subgroup (i. e. A61K6/02) as most widely used levels of the IPC hierarchy. Some IPC codes are assigned consistently and are suitable for searching using IPC codes only, while others require searching with help of key terms.

3.2.3. Participant Beta

Due to their background, this participant has some experience in text mining, especially annotating patent texts. Accordingly, initial questions led to a detailed discussion on this topic.

Participant Beta pointed out that machine translation, Optical Character Recognition (OCR) artifacts and different writing styles make automatic analysis error prone. Participant Beta repeated Participant Alpha's assertion about quality of the data playing a crucial role in automated analysis. The participant elaborated that sometimes the user might get an impression of some trends happening, while in reality they only can be attributed to the noise and errors in the data.

According to Participant Beta, it is difficult to separate citations from the rest of the text, and noise in data occurs when this process was not successful. Nevertheless, assuming that citations were recognized successfully, they constitute a very important kind of connection between documents and represent a very strong similarity. If A cites B, but they do not have similar key terms, the approach might be faulty. Patent families should be grouped in an obvious way as well.

Considering usage scenarios, Participant Beta reported that in 80% of all cases solely the distribution of assignees plotted by publication year was sufficient. Moreover, participant expressed doubt about value of visual patent landscaping tools compared to traditional non-visual tools. The participant themselves as well as other patent experts have difficulty interpreting the visual representation of a landscape. The connection between terms that

describe clusters is perceived as far from obvious and needs an explanation. The meaning of valleys between the mountains in tools such as STN AnaVist is also confusing. It is unclear to patent experts how one is supposed to recognize patterns and draw conclusions from such representation and what additional knowledge it provides. Dislike for black-box algorithms was expressed by the participant and their acquaintances.

The participant provided some advice concerning processing of the textual content.

First, they recommended using patent title and abstract together because the title itself might be too short (i. e. one word) or not expressive enough.

Second, they advised to pay special attention to stopword removal. There are stopwords that apply for all English texts, such as “is”, “have”, etc. Stopwords specific to patent domain include structural markers such as “patent application”, “description”, “prior art” or“claim”. The participant reported that during demonstrations of patent analysis software, they have regularly seen stopwords which should have been removed.

Third, Participant Beta highlighted the importance of not only single-word key terms but phrases that are two or three words long. Such phrases may consist of nouns but also of adjectives and participles. However, the participant’s impression was that single-word key terms do not appear sufficiently often after key term extraction. The participant disapproved of this, because some domains possess highly relevant one-word key terms. The situation was attributed by the participant to insufficient weighting by the extraction algorithm.

Participant Beta stated that the description field in a patent can be separated into multiple segments such as “Background of the Invention”, “Summary of the Invention”, “Brief Description of the Drawings”, “Detailed Description of the Preferred Embodiment” and “Claims”. The segmentation is not a straightforward process because the format of the parts of the description varies significantly. Assuming the segmentation is successful, “Summary of the Invention” and “Detailed Description of the Preferred Embodiment” provide most value. The participant expressed curiosity to compare results of a semantic approach between the above-named description segments and “Background of the Invention”.

3.2.4. Participant Gamma

Participant Gamma had most experience with very specific questions from clients that needed quite specific answers. Typical size of a query result for them is 30-50 patents from which about 5 most relevant have to be selected and presented to the client.

Participant Gamma repeated statements from Participant Alpha about the iterative nature of a patent search. They both described the process in a very similar way: one should approach a search from multiple perspectives, develop multiple strategies and merge them at the end.

Participant Gamma explained that freedom-to-operate research needs very complete answers to the search query. It is very important that no relevant patents are missing from the dataset, otherwise absence of any already protected inventions cannot be proven.

When it comes to working with single patent documents, the workflow is as follows. First title and abstract are read. When those are interesting enough, the full text is requested (which might require an additional fee) and studied.

Participant Gamma echoed Participant Alpha's sentiment about limited usefullness of IPC classes. For broad analyses they are found helpful, while the more specific the search becomes, the more weaknesses IPC shows.

Due to the very narrow nature of requests Participant Gamma works on, they did not see data visualization as the main tool for the search task for them personally. Instead, they saw it as a way to generate impressive visuals for the stakeholders or other non-experts. Nevertheless, they admit that visualization may be useful when dealing with large datasets. In this case, interaction and usable filtering features are seen as very important. Participant Gamma acknowledged that demand for such patent visualization tools exists among their clients.

3.2.5. Findings from user interviews and their implications

In this section, we briefly summarize the findings from the user interviews and how they influence our approach:

- Thematic consistency of IPC classes declines as one moves to the lower levels of IPC hierarchy. It also varies a lot depending on the specific technology domain. One therefore should not expect a clear separation of classes in the semantic space during the clustering (see subsection 5.1.7 for details) and the evaluation (see chapter 6 for details on evaluation). Nevertheless, patent experts work with all levels of IPC hierarchy.
- Names of assignees (companies or individuals a patent belongs to) are often spelled differently throughout the dataset. This makes grouping patents by assignee name unreliable. Disambiguated assignee names are one of the characteristics of a quality dataset. We therefore merge assignees with similar names via fuzzy string matching as a part of the preprocessing before we aggregate the documents (see subsection 5.1.3 for details).
- Patent language differs in form and structure from the common written language. Both general and very specific vocabulary is used in patent texts, and, accordingly, in patent searches. Consequently, we use a language model trained specifically on patent texts because it reflects the peculiarities of the patent domain best. We then attempt to produce generalizations of cluster key terms as described in subsection 5.1.7.
- Special attention has to be paid to the elimination of stopwords that add no value to the general understanding. We describe the process of the stopword removal and the stopword lists we use in subsection 5.1.3 "Stopword removal".
- Patent's title is rarely sufficient to describe its content and thus should be considered together with the abstract. The "Description" field of a patent consists of segments

that vary in relevance and are difficult to separate. Since our data source does not provide a segmented description, we use another textual field named“Claims” instead as described in subsection 5.1.3.

- In patent-specific language there are a lot of established expressions that consist of multiple words. This means it is not sufficient to only extract single-word key terms, which is why we extract both unigrams and bigrams to characterize content of patent documents and document clusters as described in subsection 5.1.5.
- Belonging to the same patent family constitutes the strongest kind of connection between patents. In most cases, patents from the same family should be adjacent in the semantic space. At the same time, families with diverging content exist where this rule of thumb does not apply. We use proximity of patent families as a minimal criterion which has to be met in the visualization space (see subsection 5.1.6 for details).

4. Visualization concept

In the previous chapter we described how the interviews with the patent experts shaped our understanding of the patent landscaping task. The knowledge gained in the process combined with the ideas gained from state-of-the-art approaches allowed us to develop a concept for the visualization. In this chapter, we first outline the visualization concept in its final state. We then cover the decisions that led to this stage. Lastly, a brief summary of the data processing steps needed to produce document representations is given. In the next chapter (chapter 5), we present the separate components of the user interface of the visualization and of the data processing pipeline in their implemented form and discuss them in detail.

4.1. Outline

In this work, we deal with semantic exploration of documents. This refers to, on the one hand, the challenge of displaying high-dimensional *semantic representations* of documents visually. On the other hand, we support *semantic interactions*, which means that the display adapts to the intentions of the user with regard to information density and level of detail. These two topics are reflected in our concept.

A simplified representation of the visualization layout can be seen in Figure 4.1. The user interface consists of four interconnected parts: scatter plot, histogram, detail view and sunburst including breadcrumbs.

The scatter plot is the main area of the visualization where each patent document is represented as a point. Visualization space within the scatter plot is, effectively, the high-dimensional semantic space reduced to two dimensions. The user can navigate this space by panning and zooming (see subsubsection 2.1.1.3 for details on panning and zooming). Each document is labeled by its relevant key terms which are extracted as described in subsection 5.1.5. To keep the point labels readable, a heuristic is applied for optimal text density depending on the number of points within view and the zoom factor (see subsubsection 5.2.1.1 “Zooming” for details). Additionally, points and labels increase in size slightly to support the feeling of “moving into” the data. Hovering over a patent makes its family connections, forward and backward citations to become visible as lines of different color and stroke type.

The documents are grouped into clusters that are characterized by a list of key terms. The three most relevant key terms per cluster are always visible, and a full list of top 15 key terms is shown on demand when the user hovers over the cluster label. Moreover,

additional context for every single-word key term is provided by a list of words which are semantically similar to the term and occur often within the cluster. We call those *augmenting terms* and describe how they are generated in subsection 5.1.7.

There are three sizes of clusters produced by a hierarchical clustering algorithm. As the user zooms into the scatter plot, large clusters are first substituted by medium-sized clusters and then by small ones. This is an embodiment of the semantic zooming (see subsubsection 2.1.1.4 for details on semantic zooming).

The histogram and sunburst views present metadata attributes from the dataset in an aggregated form. They enable filtering of the the data points within the scatter plot by brushing and linking (see subsubsection 2.1.1.5 for details on brushing and linking). When filtering happens, a subset of the points in the scatter plot becomes grayed out, so that the user can focus their attention on the remaining documents.

The histogram shows the temporal dynamics of patent activity represented by the number of patent applications per year. A user can select a time interval by brushing . A corresponding filter is then applied to the data.

The sunburst is essentially a stacked pie chart. It shows the distribution of patents across a given set of metadata attributes in the form of a hierarchy. Patent assignee, country and IPC classes can be used as levels for the sunburst hierarchy by themselves or in combination with each other. The user can navigate back and forth between the sunburst hierarchy levels. If one sector in the sunburst is clicked, it becomes the current root node and its children occupy the whole circle. Then, with a click in the middle of the sunburst the user can go one hierarchy level up. Moreover, colors of points in the scatter plot correspond to colors of sunburst sectors in their current state. Pie-chart-shaped glyphs appear in place of points where there are multiple values per document, for example, multiple assignees.

The sunburst is complemented by *breadcrumbs* analogous to the ones seen in website navigation. They show the currently selected sunburst node and its predecessors. After the breadcrumb path the percentage of all documents that correspond to the currently selected sunburst node is displayed.

The detail view offers the possibility to study one patent document thoroughly in accordance with the principle of details-on-demand (see subsubsection 2.1.1.1 for details-on-demand). The details for a patent become visible in the detail view when the patent is selected by clicking or hovered over.

All views are coordinated through user interactions, which fall into tree groups: selection, highlighting and resetting the current selection. Those are implemented in a consistent way across all views: hovering with a mouse causes a highlighting of an object/group which is a preview of the selection, clicking means selecting an object/group and clicking on the background of a view resets the selection. We apply the visual information seeking mantra (described in subsubsection 2.1.1.1) to allow efficient exploration of the data.

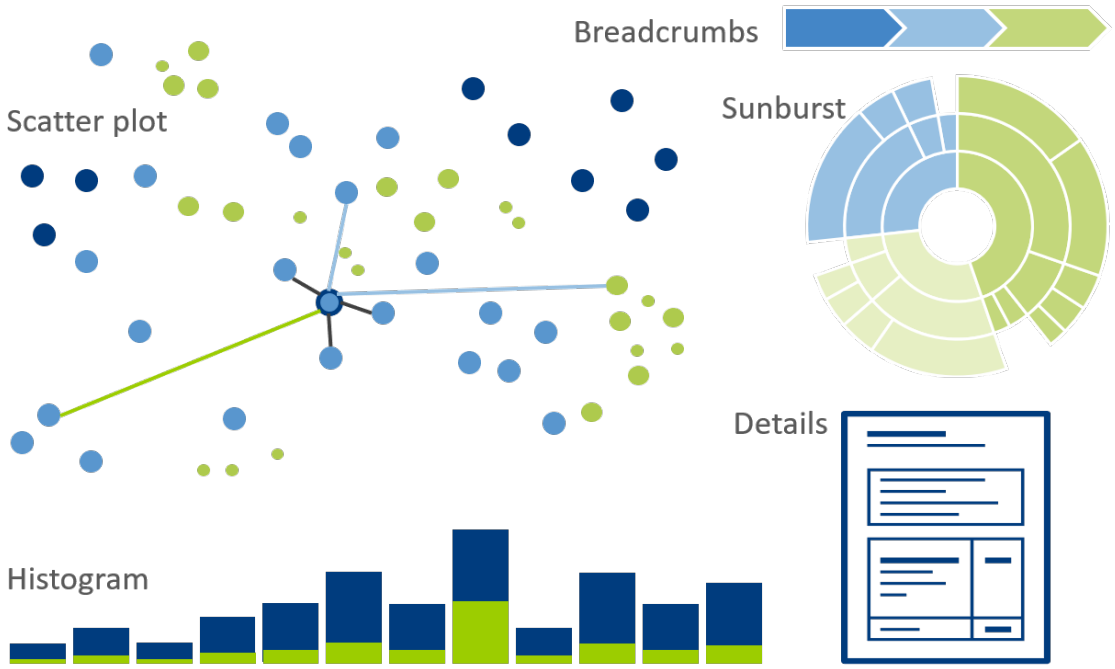


Figure 4.1.: A schematic representation of the visualization layout.

4.2. Ideation

In this section we justify the decisions that led to the creation of the visualization concept. We then discuss the first iteration of the concept and the changes it went through during the development.

4.2.1. Dimensionality of the visualization space

The dimensionality of the embeddings that we intend to visualize is much too high to be plotted as it is. A lower-dimensional presentation of the data must first be obtained via a dimension reduction technique. For that, a choice between 2D- and 3D-representation has to be made. The latter provides an advantage in the sense that one additional dimension of the data can be displayed. However, this gain comes with trade-offs with regard to usability.

Nielsen [76] discourages using 3D for user interfaces. He argues that while navigation in a 3D space looks impressive for an observer, it requires more cognitive resources from the user. First, current interaction techniques are not specifically adapted for a 3D space and are cumbersome. Second, even if the user successfully masters the controls, they still have to pay extra attention to navigating the 3D view in addition to navigating the information space. The perspective itself introduces some usability problems. For example, remote objects are often hidden by nearby objects or are too small to be readable. Additionally, it might be difficult to estimate the exact depth of an object or the distance between objects.

The plausibility of this line of reasoning has to be tested empirically. Westerman et al. [110], Banchs [10] and Fabrikant [37] performed comparison studies of 2D vs. 3D with regard to information retrieval tasks.

Westerman et al. evaluated searching for objects in semantic spaces with multiple options for the amount of variance explained by all dimensions together. They found that performance was generally poorer in three-dimensional condition with comparable amount of variance to a two-dimensional condition. Moreover, they suggest that three-dimensional interfaces "incur greater cognitive costs because of the demands of a more complex semantic mapping, i.e. maintaining a more complex mental model of the information space".

In works of both Banchs and Fabrikant, 3D interfaces received positive feedback and were the participants' preferred representation. In Banchs' study, the participants reported that the 3D platform allowed faster search, when in fact task completion times were lower for the 2D platform. Notably, a higher percentage of tasks was accomplished successfully using a 3D platform. Nevertheless, Banchs' conclusions match Nielsen's reasoning, namely that 2D interfaces are currently still more familiar to users. Banchs highlights the performance of the visualization as one of the significant limitations, which was also mentioned by the participants. This limitation also applies to our work, since our goal is to enable exploration of thousands of documents.

A notable exception from mentioned negative aspects constitute applications with entertainment purposes or for rendering physical objects in their solid form, where using 3D is encouraged [76]. Moreover, all above-mentioned arguments only apply assuming a pseudo-3D representation on a conventional two-dimensional computer screen. A Virtual Reality (VR) application would define its own interaction techniques that feel natural for a 3D space. Immersive data visualization in such environment using a VR headset has been researched, for example in [33] and [44]. Moreover, using a VR environment implies a technology stack that is better adapted to displaying complex geometry, for example large point clouds. Unfortunately, the advantages of a "true 3D" approach are not utilizable on a standard desktop workstation without extra hardware.

Ultimately, gaining one additional spacial dimension for representation is not worth the increased inconvenience of navigating the information space. Therefore, we decided upon a 2D representation for our prototype.

4.2.2. Choice of a suitable visualization metaphor for hierarchical data

Most of the patent metadata attributes are of relatively common data types like date, string or list of strings. But there is one attribute with an uncommon type and that is the IPC class (described in subsubsection 2.1.3.4), or more specifically, a list of IPC classes. IPC classes are hierarchical in nature, which necessitates a suitable visual metaphor.

The *treemap* as shown in Figure 4.2 is a space-filling type of diagram that was proposed by Shneiderman [94] and has often been used to visualize hierarchical data. It uses nested shapes, usually rectangles, to represent the parent-child relationship. A parent rectangle's space is divided into child rectangles along an axis that changes with each nesting level. Size

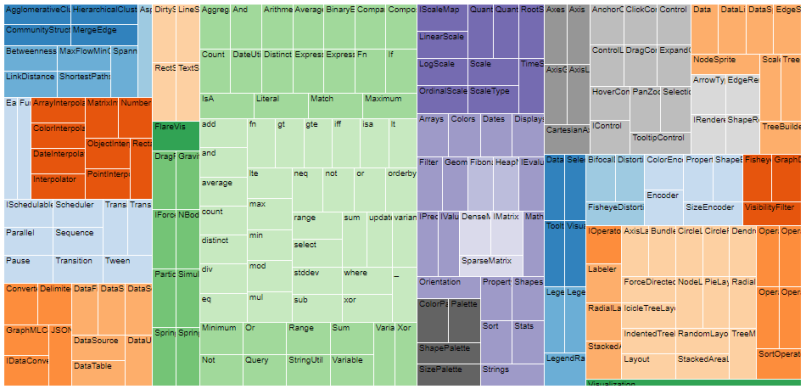


Figure 4.2.: Treemap visualization of the class structure in a programming library Flare. Source: [101]

and color of rectangles represent various attributes of hierarchy nodes. In an interactive version of a treemap, the user can select a node to examine its children in detail. The available space is then redistributed to a subset of the hierarchy with the chosen node as its top.

The *sunburst* type of diagram was inspired by the treemap. It utilizes a radial layout in which child nodes are not contained in the parent nodes, but expand outwards from the circle center. Size (angle) and color of the radial sectors can, just as with a treemap, represent chosen attributes of nodes in a hierarchy. One can also navigate within the hierarchy by choosing a node to serve as a starting point in the center.

[99] evaluated treemap and sunburst in their study. Their conclusion was that the sunburst "more frequently aided task performance, both in correctness and in time, particularly so for larger hierarchies. The explicit portrayal of structure appeared to be a primary contributor to this benefit." Supported by this finding, we initially chose to use a sunburst for a fairly large hierarchy that is the IPC classification. Later, an idea emerged that the attributes represented by sunburst do not necessarily have to be of a hierarchical nature. It is possible to "stack" multiple categorical attributes, for example country and assignee, to produce subgroups/child nodes which are represented in a sunburst. Moreover, it is also possible to combine categorical and hierarchical attributes to show, for example, a distribution of IPC classes per country. In our visualization concept, we evaluate the feasibility of using metadata attributes of various types in a sunburst diagram.

4.2.3. Initial concept and its evolution

The initial concept was inspired by a demonstration of coordinated views showing mock data by [53]. This demonstration implemented brushing and linking (see subsubsection 2.1.1.5 for more on brushing and linking). The demonstration consists of a scatter plot on the left, a histogram on the right and an additional representation of a time axis at the bottom (see Figure 4.3).

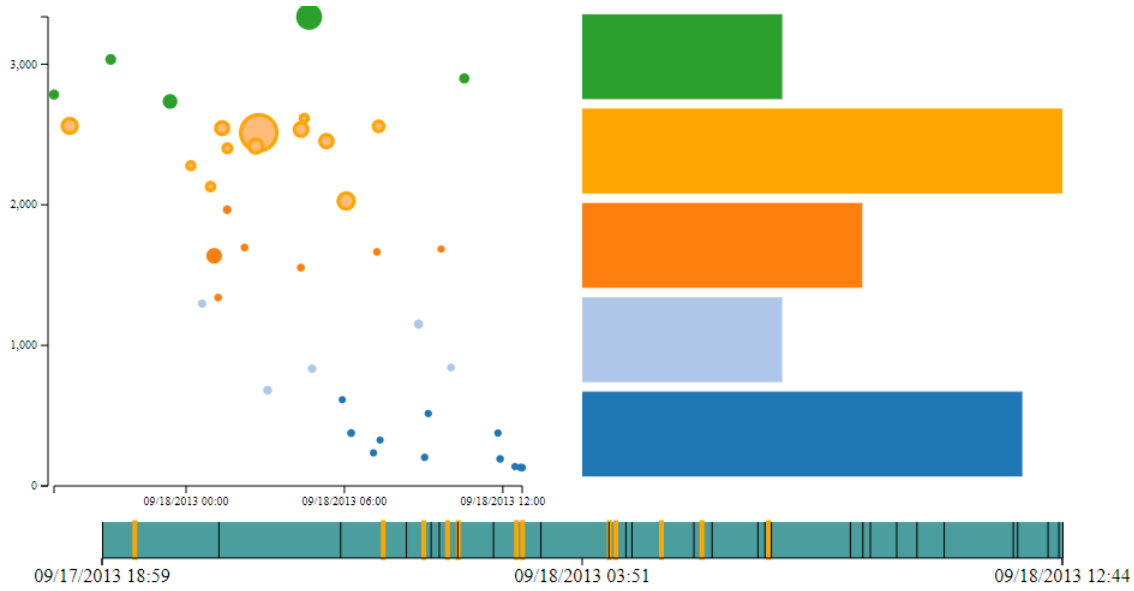


Figure 4.3.: Demonstration of coordinated views that served as an inspiration for our concept. Image source and demo: [53]

The Y-axis of the scatter plot corresponds to a numerical dimension in the mock dataset, while the X-axis represents timestamps of data points. The histogram splits the values of the numerical dimension into five bins. The element at the bottom of the demo is essentially a *rug plot*, i. e. it denotes the X-positions of data points by tick marks that look similar to tassels on a rug. A classic example of a rug plot can be seen in Figure 4.4.

Rug plots are used to illustrate the distribution of a variable along the axes, in this case the time axis. Usually, a rug plot is drawn as a part of the original plot (scatter plot, line plot, histogram, etc.), but in this instance it has been separated from the corresponding scatter plot. The main purpose of the element is to enable selection of the data by brushing and linking, which is why we and the author refer to it as the *brush element*. The histogram values are recomputed in accordance with the updated selection. Moreover, when the mouse is hovering over the histogram bins, corresponding data points are highlighted both in the scatter plot and in the brush element.

The first iteration of our visualization concept (see Figure 4.5) was built on ideas borrowed from the above-mentioned demonstration by [53].

Firstly, it was the idea of a main area that contains data points and is influenced by controls on the edge of the display. Displaying distribution of patent applications over time and the ability to select a time interval we consider especially useful for the patent landscaping use case. Therefore, we decided to implement the same linking and brushing functionality, yet considering the size of the data, a histogram with yearly bins was judged more fitting to show the distribution over time.

Secondly, inspired by the ideas from [53] and [112], our UI concept was designed to show a distribution of the dimensions of the data. In our case those attributes are mostly categorical, and IPC classes are also hierarchical in nature. The wish to display multiple dimensions of

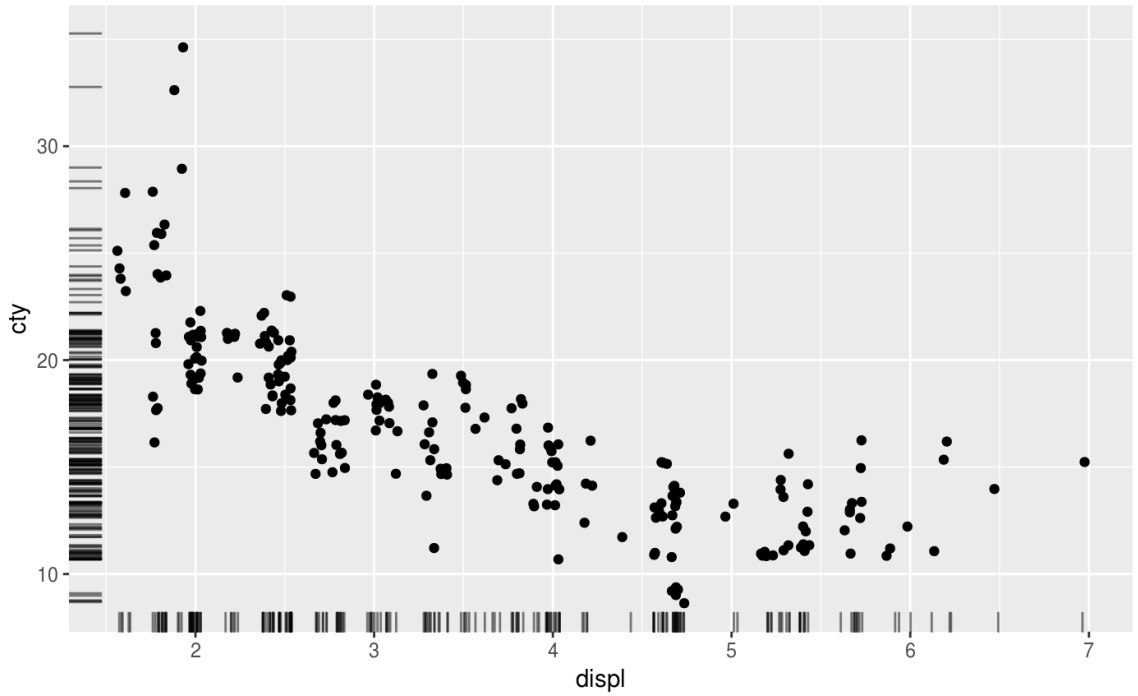


Figure 4.4.: A scatter plot is augmented with a rug plot. The rug plot shows distribution of the data points with regard to X and Y coordinates. Image source: [89]

the metadata resulted in our choice of a sunburst chart in the place of a histogram in the original demonstration. This decision was partly motivated by Wittenburg et al. [112], who make extensive use of metadata in their faceted visualization (see Figure 2.8). They show the assignee, country and application year as a vertical stack of blocks where the width of a block corresponds to the number of patents with the corresponding attribute value. Unfortunately, their approach results in a cluttered view and therefore lacks visual scalability. We address the scalability problem via interactivity, i. e. through the fact that it is possible to change levels of a sunburst chart 1) through navigating up and down the hierarchy of attributes and 2) by choice of different sets of metadata attributes to be charted. For more details on our implementation of the sunburst chart see subsection 5.2.3.

The main area of our visualization was initially conceived as a fully connected graph. Similarities between each pair of documents were supposed to correspond to the attraction forces in the force-directed graph layout. When a part of the dataset would be eliminated through filtering, the corresponding graph nodes would disappear and the whole layout would rearrange itself. Effectively, the process would amount to computing and then dynamically updating a t-SNE representation of the dataset. An interactive demonstration of such a layout is presented in [100]. While it would certainly be of value to cluster subsets of the data dynamically depending on the selection, performance considerations outweigh the benefits. Therefore, a scatter plot with static positions of data points was chosen as a viable alternative.

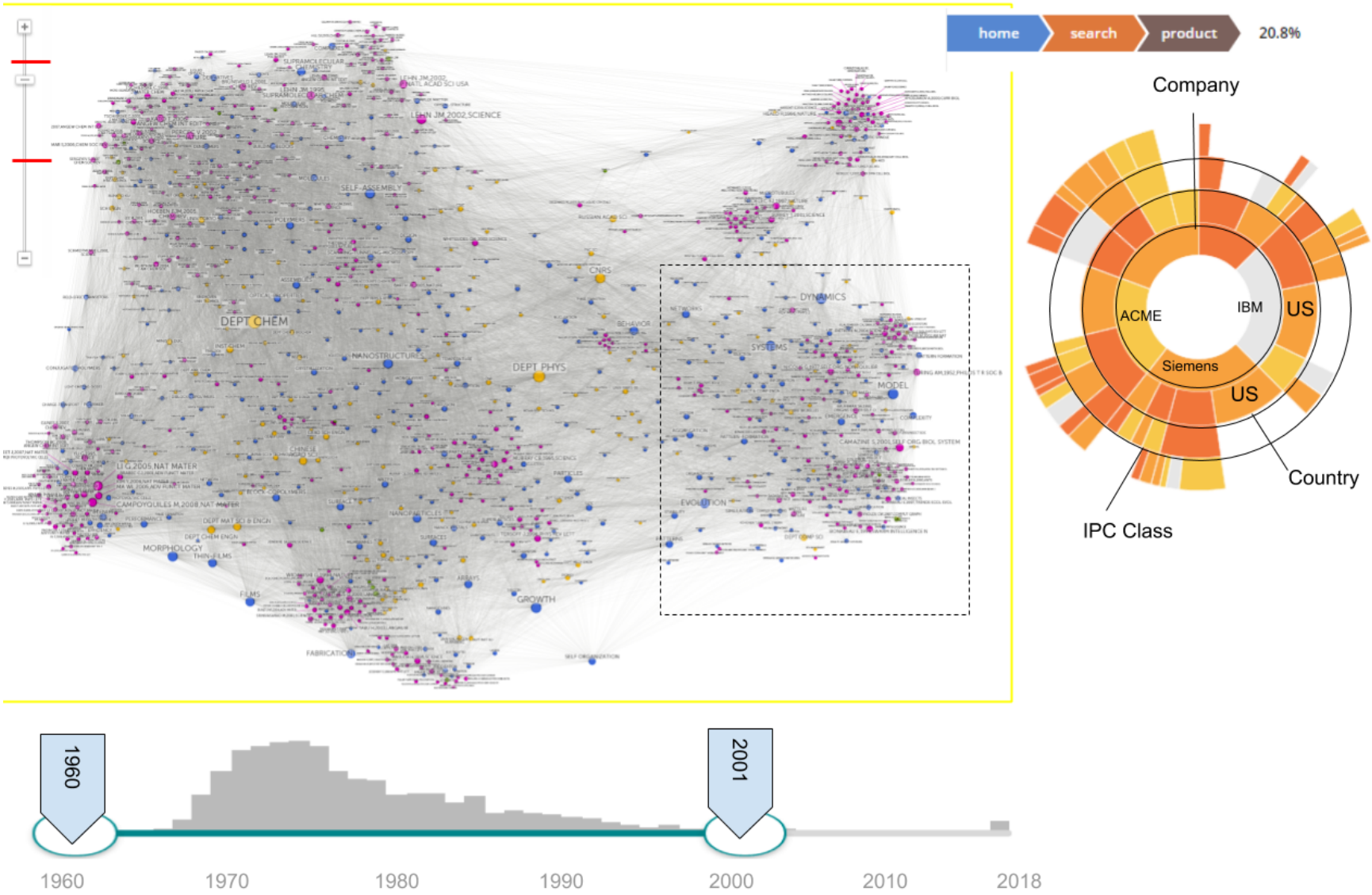


Figure 4.5.: The first iteration of the visualization concept. Source of the graph picture: [58], source of the sunburst picture: [86].

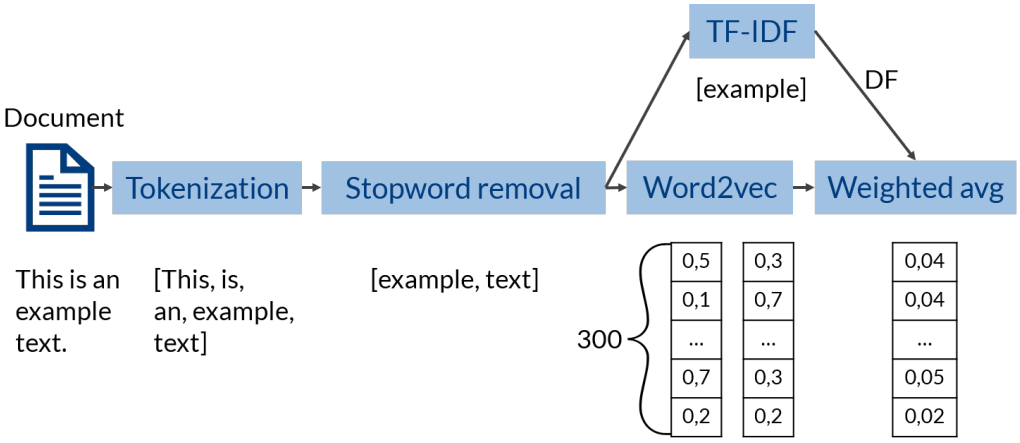


Figure 4.6.: Pipeline of processing steps for a single patent document

Initially, the concept included no standalone detail view. Instead, the idea was to display some tooltip elements directly above the currently chosen patent and above patents related to it. Type and amount of the information presented in the tooltips were supposed to change depending on current selection and zoom level according to the principle of *semantic zooming*. Upon further consideration it became clear that such tooltips would cover a significant portion of the scatter plot and would therefore render it unusable. The decision was made to place the detailed patent information to the available space in the lower-right corner.

The zoom control in the upper-left corner is an idea borrowed from various interactive map interfaces. The red markings were supposed to show boundaries between different detail levels of hierarchical clustering. Ultimately, we implemented different visual indications that sufficiently support the feeling of "moving into" the landscape and back, so this part of the initial concept was omitted. Other elements were utilized in the prototype without changing much.

4.3. Data processing

Before a dataset can be displayed in a visualization, it has to be processed in a preparatory step. Figure 4.6 and Figure 4.7 present an overview of the processing pipeline which is necessary to produce data for the visualization. In this section, we give a brief overview of the steps, which are covered in more detail in section 5.1.

First, each patent needs to be processed individually (see Figure 4.6). This starts with splitting the textual part into separate words and removing stopwords. Stopwords include general grammar-related words such as "is", but also patent-specific vocabulary such as "embodiment". Then, for each word a 300-dimensional embedding is retrieved from the word2vec model explained in more detail in subsection 5.1.1. A vector representing the whole document is composed by aggregating the word vectors as a weighed average. Each word is weighted with its IDF. Our purpose is to make semantic similarities and differences

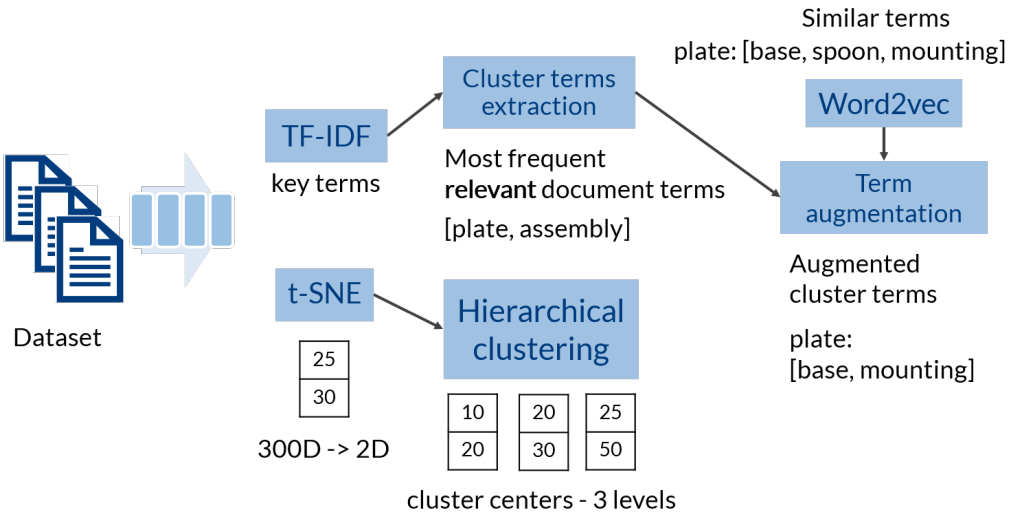


Figure 4.7.: Continuation of the pipeline after all individual patents have been processed

tangible, but numerical document vectors do not provide an explanation of why any pair of documents are close or distant. Therefore, TF-IDF is also used to extract relevant key terms per patent.

Second, after all document vectors and document key terms are computed, the processing on the dataset as a whole can begin (see Figure 4.7). 300-dimensional document vectors are reduced to two dimensions with t-SNE to fit the visualization space. This enables hierarchical clustering, which splits the dataset into a number of clusters on three detail levels, resulting in large, medium and small clusters. It is crucial for user's understanding to know what common semantic characteristics the grouped patents share. To cover that, we extract the key terms per cluster. We take most relevant terms per document and count them across all patents within a cluster. The terms with most occurrences are assigned to the cluster to explain its thematic focus.

To aid the understanding of cluster key terms, for each single-word term we retrieve similar words from the above-mentioned word2vec model. We then check whether those similar words occur in more than 10% of the patents in the cluster. If that is the case, the word is added to the list of augmenting terms for the given key term. For example, if we try to augment the term *plate*, the word2vec model might retrieve *base, spoon, mounting* as similar words because they often appear in similar contexts with *plate*. Assuming the dataset is not about cutlery, *spoon* does not appear in many patents, but *base* and *mounting* might. Therefore, *base* and *mounting* are the terms that provide context for the term *plate* in the given cluster.

Herewith the processing of the textual part of the patent dataset is complete. As for metadata, some of the attributes can be used in the visualization as is, while others require special processing (see subsection 5.1.3 "Parsing of metadata attributes" and subsection 5.1.4). We elaborate on the processing of both textual and metadata parts of a patent dataset in section 5.1.

5. Implementation

After having defined the concept for our approach as described in the previous chapter, we created a proof-of-concept prototype to evaluate our approach. In this chapter, we describe both the visual part of the implementation and the behind-the-scenes processing which serves to prepare the data to be visualized.

The code of the implementation can be found at `https://github.com/gingergenius/patent-embedding-visualization`.

5.1. Implementation of the data processing

In this section we describe what data source we use. We then justify our choice of programming languages and tools. Finally, we elaborate on the data processing pipeline described briefly in section 4.3.

5.1.1. Data source

Google Patents Public Datasets [111] is a data source available for public use on Google's BigQuery platform. It contains full-text patent publications for the US and bibliographic data (abstract + metadata) for patents filed in the rest of the world (see subsubsection 2.1.3.1 for the description of data fields per patent). The database has about 100 million patent document records and is updated quarterly. While a quick examination of a sample of the data revealed some format errors, probably resulting from optical recognition, most of the text is readable. This data source is therefore sufficient for a quality analysis.

We extracted four separate datasets about different technology domains from Google Patent Public Datasets:

- *Hair dryer* contains approximately 250 patents. It was included in the codebase from [3] and consists purely of patent applications from the US. Because of relatively small size of this dataset, separate thematic areas within it are not clearly distinguishable.
- *Video codec* contains about 1600 patents. It was also included in the codebase from [3] and consists purely of patent applications from the US. This dataset consists overwhelmingly of patent families of different sizes. Because of that, it is suitable for evaluating how well the family similarities are handled by the semantic embeddings. However, because of our unfamiliarity with the topic of video encoding, an alternative dataset was necessary for the evaluation.

- *3D printer* with roughly 1000 patents was produced by us using a query adapted from an existing patent landscaping report [1]. The full text of the query can be seen in section A.2. The dataset consists of both US and non-US patents. Non-US patents have only abstract text available while US patents have all textual fields available, of which we use the abstract and claims. We wished, however, to exclude effects resulting from different text lengths, but still make a considerable number of patents available for exploration. Consequently, we additionally prepared the contact lens dataset.
- *Contact lens* consists of ca. 2600 patents and was based on a query adapted from an existing patent landscaping report [29]. The full text of the query can be found in section A.3. We restricted this dataset to US-only patents so that we would be able to work with both the abstract and claims for the whole dataset. The contact lens dataset contains a variety of topics and is large enough for interesting exploration tasks. This is why we chose it for our evaluation.

Additionally, FIZ Karlsruhe kindly provided one more dataset extracted from the World Intellectual Property Organization (WIPO) database on the topic of diesel engines (ca. 4500 patents). The patents in it had both abstract and claims available, but no citation or family information. Moreover, multiple languages were present in the text and everything beside English had to be filtered out. This dataset was necessary to check how the approach performs on large data. Unfortunately, it could not be used for the evaluation with experts because of the missing metadata attributes.

Many existing pretrained models for word embeddings are based on general vocabulary. Language and especially vocabulary in patent documents deviate significantly from general speech, which must be taken into account. Abood et. al [3] provide a word2vec model trained on 5.9 million patent documents. It contains a 300-dimensional embedding for each of 110239 words in its vocabulary. We make use of this model for our semantic analysis to compute document embeddings.

5.1.2. Choice of technology

Python [83] is chosen as the programming language for preparing the data for visualization. It is an extremely widespread language in the field of machine learning with a great number of libraries available. Of those libraries, Tensorflow [103] in combination with Keras as a high-level Application Programming Interface (API) [55] is a state-of-the-art library for neural networks. In fact, [3] used them in their approach for data preprocessing and for creating the word2vec model we use. We based our approach on their codebase, so Tensorflow and Keras were also inherited by us. Other widely used Python libraries we take advantage of are Numpy [79] and SciPy [92] for scientific computing and scikit-learn for machine learning. Finally, JupyterLab [54] with an IPython [50] kernel is selected as the development environment of choice because of ease of quick prototyping.

For the interactive visualization itself, information visualization frameworks such as Axiis [9], Bokeh [15], D3.js [17], Altair [6] and several others were considered. For the given problem, creating both custom user interaction techniques and custom visualization

layouts consisting of new forms of charts is required. Unfortunately, most visualization frameworks do not support this. Instead, they restrict the developer's alternatives to predetermined chart types. Coordinated interactions between views are either not supported or very limited. Consequently, D3.js is chosen as the visualization framework with most flexibility. It requires a hands-on, low-level approach to programming, where the developer has to manually define SVG shapes and bind their attributes such as position or color to the data. However, this is exactly why D3.js provides the necessary level of control for the implementation of our prototype.

5.1.3. Data preprocessing

In subsection 5.1.1, we mentioned multiple datasets that were prepared for the visualization in the course of this work and the queries used to produce them. In this section, we describe the preprocessing steps that are executed for each patent document within a retrieved dataset.

The diesel engine dataset was, unlike the others, not derived from Google Patents Public Datasets, so it required some minor additions to the pipeline. The data structure was slightly different and therefore had to be transformed for compatibility. Moreover, the patent texts had to be cleaned as there were some Extensible Markup Language (XML) tags present that we removed. Additionally, the diesel engine dataset included a non-negligible amount of text in French and German within the patent claims. To filter out non-English text, we split the text into sentences and detected their language using the Python library langid [68]. This language detection tool utilizes a multinomial naive Bayes classifier trained on n-grams to reliably produce a robust result independently of domain and text length [67].

The interviews with the patent experts indicated that title, abstract and claims are the most valuable textual parts of a patent document for understanding the described invention. Other textual parts, such as background art or description of figures, do not contribute significantly to the essence of the invention. Therefore, for all further steps we concatenate the patent's title, abstract and, when available, its claims. This way, we maximize the length of relevant textual content taken into account. We do not consider the three above-mentioned textual fields separately for sake of simplicity. However, for future work it might be worth examining how different textual parts of the same document compare to each other when semantic methods are applied to them separately.

Tokenization

The proper preprocessing starts with *tokenization*, which means splitting the text into tokens. Tokens in our case are not characters or sentences but words since we use word embeddings from a pre-trained word2vec model. We use a `Tokenizer` class from Keras which replaces all punctuation except the apostrophe character with spaces. It then translates the whole text to lowercase and splits the text into words divided by spaces. The last step is replacing the numbers that occur separately or as parts of a word with a `_NUMBER_` token.

Stopword removal

After we have successfully tokenized the data, the next step in the pipeline is the stopword removal to increase the amount of meaningful information per document. We use two stopword lists: one is a general list of English stopwords from Natural Language Toolkit [13] and the other one is a patent-specific list kindly provided by FIZ Karlsruhe. The latter list includes words like "comprised", "abovedescribed' and "obtained" which often appear in patent texts and do not contribute to the meaning. Words shorter than 3 characters or longer than 50 characters are eliminated as well. Finally, when the number of meaningful words per patent becomes clear, we remove all patent documents that contain less than 30 words. The amount of the remaining text at this stage varies between the datasets, but in all cases it followed a left-skewed distribution with a mode of approximately 250 words and an average of 500 up to 1000 words. Distributions for diesel engine and contact lens can be found in Figure A.3.

Parsing of metadata attributes

This step is independent of the processing of textual content, but instead prepares the metadata attributes for visualization. Metadata attributes such as references, assignees and IPC classes are included in the data as a string made up of comma-separated values, so we split the list to get each separate value. For IPC classes, we compute a list of unique codes for all levels of the IPC hierarchy per document (see Table 2.1 for description of levels).

The assignee names present a challenge with regard to the data quality. Institutions' legal names are written out in a very inconsistent way throughout the data. There are often multiple variants with parts of the name which are abbreviated in some cases but not in others. Different branches or subsidiaries of the same company are also often present. Lastly, there are errors and misspellings as well, partly as a result of OCR artifacts. This last point applies to assignee names signifying private persons as well as companies.

We would like to be able to reliably group patents by their assignees. For this, we merge similar assignee names with *fuzzy string matching*, which is a technique of finding strings that are approximately the same. We use the Python library `FuzzyWuzzy` [28], which is based on Levenshtein distance between two sequences of characters. The Levenshtein distance is a measure of similarity composed of the number of character deletions, insertions or substitutions required to transform one string into another [64]. The value returned by the library is not the absolute distance but a similarity percentage that takes string length into account. We use a simple similarity threshold of 88% to determine which assignees to combine into one entry. The threshold value was chosen empirically to provide sufficiently good results. Fuzzy string matching allows us to reduce the number of unique assignees in the dataset by 10-15%. However, false positives (match detected where none exists) and false negatives (existing match not detected) could not be completely excluded from among the matches.

5.1.4. Sunburst hierarchies

As we would like to use the sunburst control for arbitrary combinations of metadata attributes, corresponding hierarchical aggregations need to be computed. For that, we consider IPC class, country, assignee by themselves and also all possible permutations (orderings) of those attributes of size two. We do not consider stacking all three attributes as sunburst levels because of space restrictions and because the groups after the third division would become very small. However, it would only be effective for larger datasets and documents with multiple non-hierarchical metadata attributes.

For metadata attributes, we distinguish between a single value per document (e. g. country), a list of values per document (e. g. assignee) and a hierarchical code such as an IPC code or a list of such codes. This allows us to adjust how the aggregation is computed depending on the type of the attribute. For *value attributes*, we can just group all documents by their unique values of the corresponding attribute. For *list attributes*, a single document can be referenced in multiple hierarchy nodes, so it should be counted multiple times. *Code attributes* are essentially multiple list attributes in a certain order (section, class, subclass, group, subgroup) with one extra condition: a subclass from one IPC code (for example, N04N) can only count as a child of its own class (N04) and not some other class (B02).

As mentioned in subsubsection 5.2.3.1, because patents simultaneously belong to multiple nodes, a total number of patents in the hierarchy may exceed the size of the dataset. In this case, the values of all nodes are normalized so that they yield 100% when combined. The normalization starts from the shallowest hierarchy level and then goes into the depth.

5.1.5. Key term extraction

We extract relevant key terms per document so that the user can get a first impression of the content of a document with just a quick glance. For this extraction we use TF-IDF which is a widely used weighting technique to determine most important words or phrases in a corpus. Essentially, the more often a phrase occurs in a document, the most important it is for this document (Term Frequency (TF)). At the same time, the more often the same phrase occurs throughout the whole document corpus, the less explanatory power it has (IDF).

For our data, unigrams (single words) and bigrams (two-word phrases) produce most meaningful results. To exclude extremely rare terms and spelling mistakes, only phrases that occur in more than ten documents are considered for their relevancy. Additionally, we also explicitly exclude terms appearing in over 20% of the corpus from consideration because they are unlikely to result in an information gain. For each patent, we save a list of ten terms that were most highly ranked by TF-IDF for visualization.

5.1.6. Embeddings

Document vectors

The purpose of this stage is to produce a representation of a document based on the words

it consists of. The input of this step is a sanitized list of words per patent, which is a result of the tokenization and stopword filtering operations described in subsection 5.1.3. We compute each document embedding as a weighed average of the embeddings of words the document contains. 300-dimensional word embeddings are retrieved from the pre-trained word2vec model provided by [3]. The weighting factor for each word is its IDF, which at this point had already been computed as described in subsection 5.1.5. Adjusting the weight of a word by its frequency in the corpus takes the varying importance of separate words into consideration and therefore helps capture themes in the dataset in a better way. Weighting with IDF has been successfully used for computing semantic similarity [114] [74], [7] and for sentiment analysis [30]. In our case, we found that compared to non-weighted word averages, weighted word vectors result in a clearer separation of clusters after dimension reduction compared to non-weighted word vectors. The documents were more likely to gather into dense groups instead of being distributed uniformly. A comparison of weighted and non-weighted word averages can be seen in Figure 5.1 and Figure A.2.

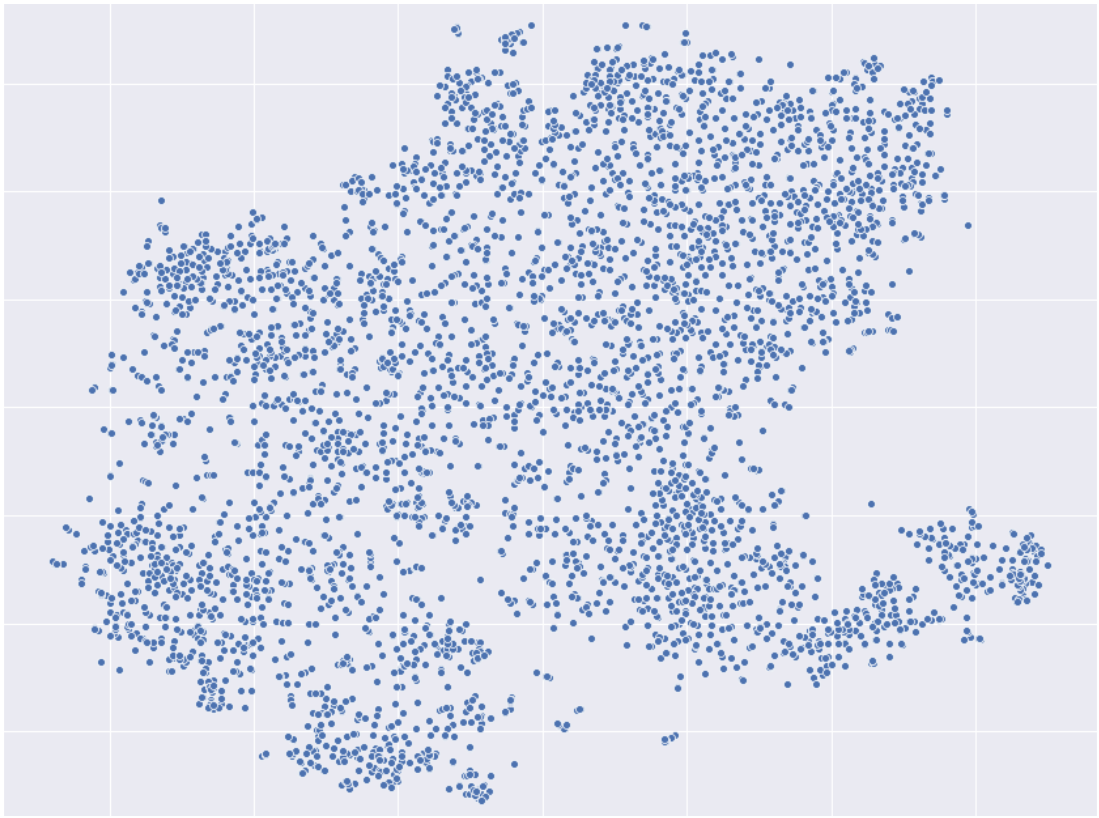
Dimension reduction

Points in a very high-dimensional space are not suitable for an understandable visualization. For the purposes of visualization we need to transform document vectors into a two-dimensional space so that any patterns in the data become recognizable.

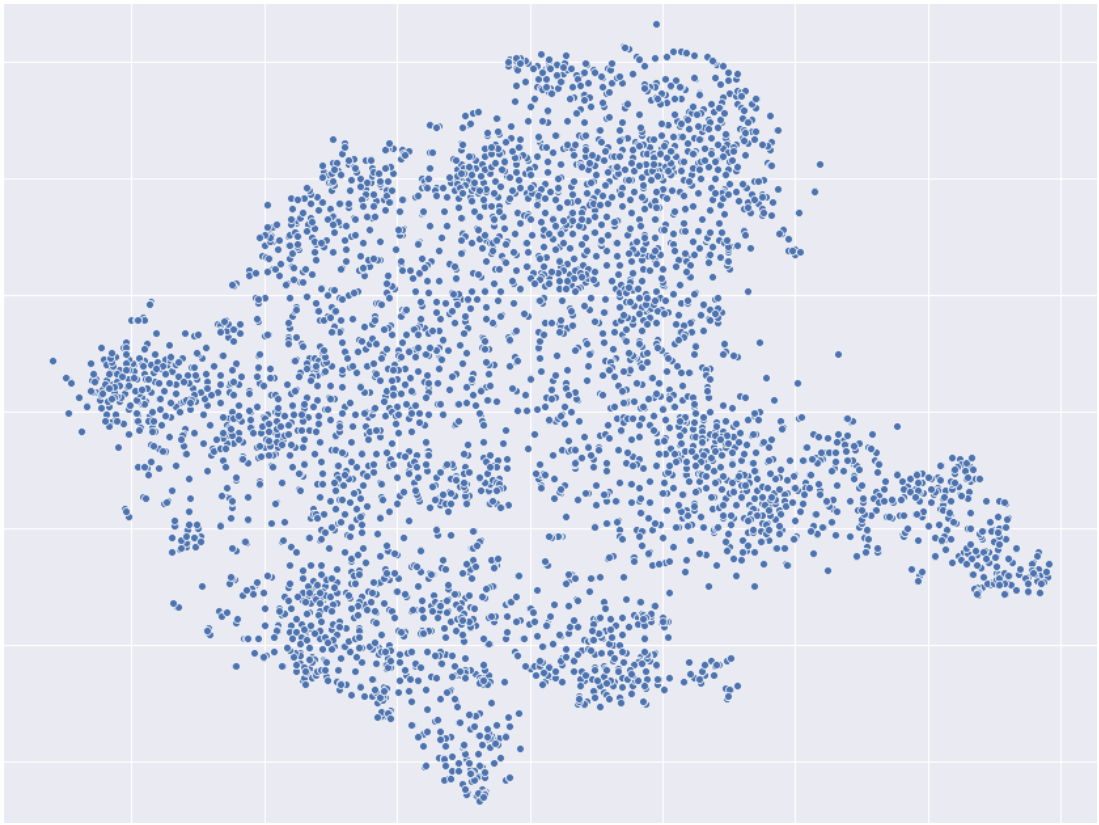
We compared multiple dimension reduction techniques such as metric and non-metric MDS [80], Isomap [102], Principal Component Analysis (PCA) [48], Uniform Manifold Approximation and Projection (UMAP) [69] and t-SNE (see subsubsection 2.1.2.2). In the interviews, patent experts emphasized that patents from the same family possess a great degree of semantic similarity and one should expect them to be placed closely to each other (see subsection 3.2.5). Handling patent families correctly is a minimum requirement for a suitable dimension reduction technique. For this reason, the comparison was conducted on the video codec dataset since it chiefly consists of patent families of different sizes.

Among the tested dimension reduction techniques, t-SNE was the only one in which families were clearly identifiable and separated from their surroundings. Figure 5.2(a) shows numerous “clumps” of closely situated points. Further inspection showed that they mostly belonged to the same family, even when the family information present in the dataset was incomplete and did not explicitly list a connection. In the majority of other cases, the patents within the groups belonged to the same assignee, dealt with the same invention and were therefore textually very similar (see Figure Figure 5.2). When tested with other datasets, t-SNE resulted in easily identifiable accumulations of points distinctly separated from each other by empty areas. Other techniques were apt to clump the points into one big area or distribute them uniformly without defined groups. Large sparse zones consisting purely of apparent outliers were also likely to appear. For comparison, see Figure A.1 for the results from other dimension reduction methods.

As t-SNE tries to retain distances from the high-dimensional space in the lower-dimensional representation (see subsubsection 2.1.2.2 for details), a suitable *distance metric* has to be used. Distances in high-dimensional spaces behave in a non-intuitive way and currently

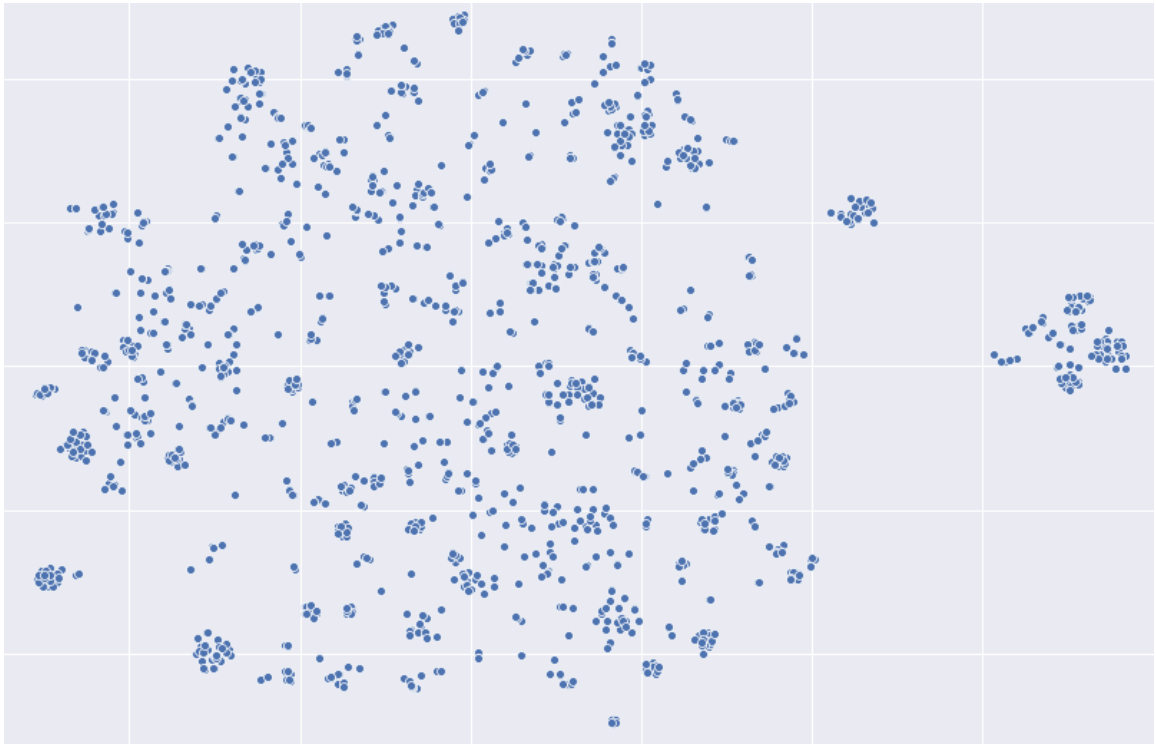


(a) Non-weighted average

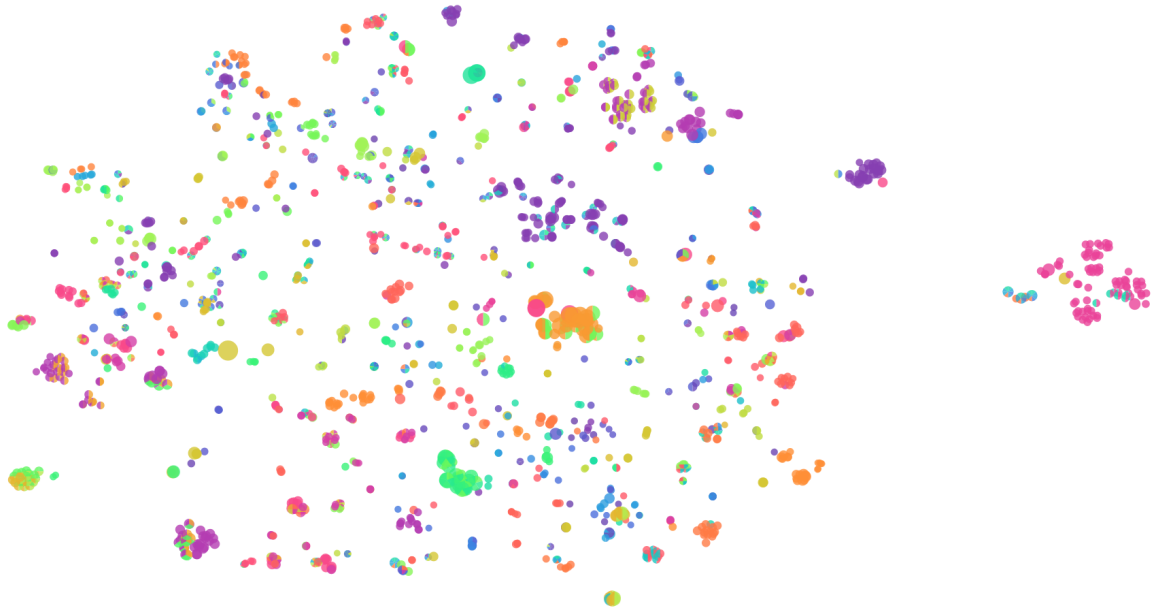


(b) Weighted average

Figure 5.1.: A comparison of document vectors computed with and without IDF weighting. Diesel engine dataset.



(a) The result of the dimension reduction by t-SNE. The points are plotted in the same color and size to make close groups visible



(b) Same coordinates as above, displayed in the interactive prototype. The patents from the same assignee are drawn in the same color, which shows that groups constitute patent families

Figure 5.2.: The result of dimension reduction by t-SNE on the video codec dataset.

there is no consensus on the "best" metric for all possible applications. We experimented with multiple distance metrics such as euclidean distance, cosine similarity and manhattan distance. The local structure of the data seemed stable independently of the used metric, only the relative placement of larger groups changed. Ultimately, we settled on *cosine similarity*, which is widely used in NLP applications. For two vectors A and B, cosine similarity is measured as the cosine of the angle between them. It can be derived easily using a dot product as shown in Equation 5.1.

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\|\|\mathbf{B}\| \cos\theta \tag{5.1}$$

As a result of this stage, 300-dimensional document vectors have been transformed into two dimensions for visualization using t-SNE with cosine similarity as a distance metric.

5.1.7. Hierarchical clustering

At this step we have 2D coordinates of all patents, but it is not immediately clear to the user why they are placed in a certain way. To explain the semantic similarities within groups at multiple levels of detail, we use agglomerative (bottom-up) *hierarchical clustering*. Essentially, is a process in which every data point is considered its own cluster at the beginning. Those singular clusters are then merged into their nearest clusters one-by-one, and in the following iterations, clusters join the adjacent clusters until the whole dataset is joined into one single cluster. The changes are tracked throughout the algorithm within a *distance matrix*, in which pairwise distances between any two clusters are stored. The process constructs a tree called *dendrogram* which reflects the structure present in the distance matrix.

An example dendrogram is shown in Figure 5.3. Points E and F are the nearest pair of points in the dataset, so they are combined to a cluster EF on the first iteration of the algorithm. The same thing happens to A and B on the second iteration. Point D and subsequently point C join cluster EF and finally, cluster AB and cluster CDEF are merged to create a root node of the hierarchy. Since it is a tree structure, there is no single correct number of clusters in a hierarchical clustering. After every merge one can decide to make a "cut" as shown by the orange line. At this specific level of detail, the dataset is then divided into a number of clusters equal to the number of dendrogram lines the cut crosses.

A distance between any two points is clearly defined in a 2D space, but multiple definitions exist for distance between two clusters. SciPy's `linkage` method, which we used as an implementation of the hierarchical clustering, offers various options for calculating the distance between two clusters:

- `single` (Nearest Point Algorithm) uses the minimal distance between points from different clusters
- `complete` (Farthest Point Algorithm) uses the maximal distance between points from different clusters

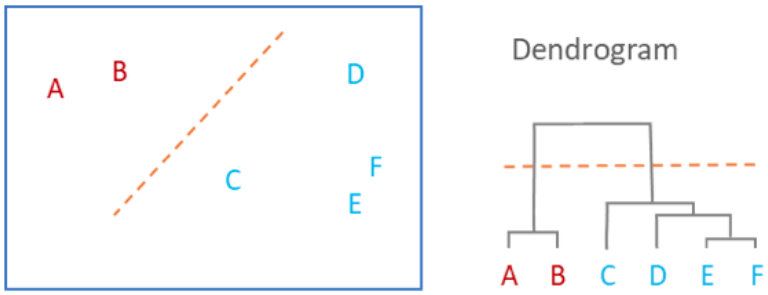


Figure 5.3.: An example of dendrogram used in hierarchical clustering and its input data. Image source: [14]

- `average` uses $d(u,v) = \sum_{ij} \frac{d(u[i],v[j])}{(|u|*|v|)}$ where u and v are the two clusters and |u| and |v| are their respective cardinalities
- `weighted` uses $d(u,v) = (dist(s,v) + dist(t,v))/2$ where clusters s and t were previously combined to form u and v is the remaining cluster

To identify the best algorithm, we computed a Cophenetic Correlation Coefficient [84] for all above-mentioned algorithms on all five of our datasets. The coefficient compares (correlates) the actual pairwise distances of all data points to those implied by the hierarchical clustering. The closer the value is to 1, the better the clustering preserves the original distances. The method `average` consistently produced higher values of Cophenetic Correlation Coefficient across the datasets, which made it our preferred method.

Besides clustering in 2D space after the dimension reduction, we experimented with clustering in the original 300-dimensional document space as well. The resulting structures were not preserved well during the dimension reduction. The clusters were not clearly divided, which means it was impossible to draw a clear boundary between clusters. This led to problems with placing cluster labels as described in subsubsection 5.2.1.2. We therefore prefer clustering data in the same space where it is visualized. This is due to a compromise that has to be made between representing the structures in the original high-dimensional space accurately and keeping the end result sufficiently simple for human cognition and therefore interpretable.

An example of a resulting dendrogram is shown in Figure 5.4. We manually chose three levels of detail for each dataset according to consistent principles. We refer to the levels of detail in terms of *large*, *medium* and *small* clusters.

- The number of large clusters should be between 3 and 7 depending on the structure of the dataset. At this level the most general topics in the dataset should be visible.
- The number of medium clusters should be between 10 and 20, so that every large cluster is divided into approximately 3 to 4 smaller topics.
- The number of small clusters should be between 40 and 70, so that every medium cluster is divided into 3 to 4 parts. This finest level of detail is aimed at summing up patent families and very closely related inventions.

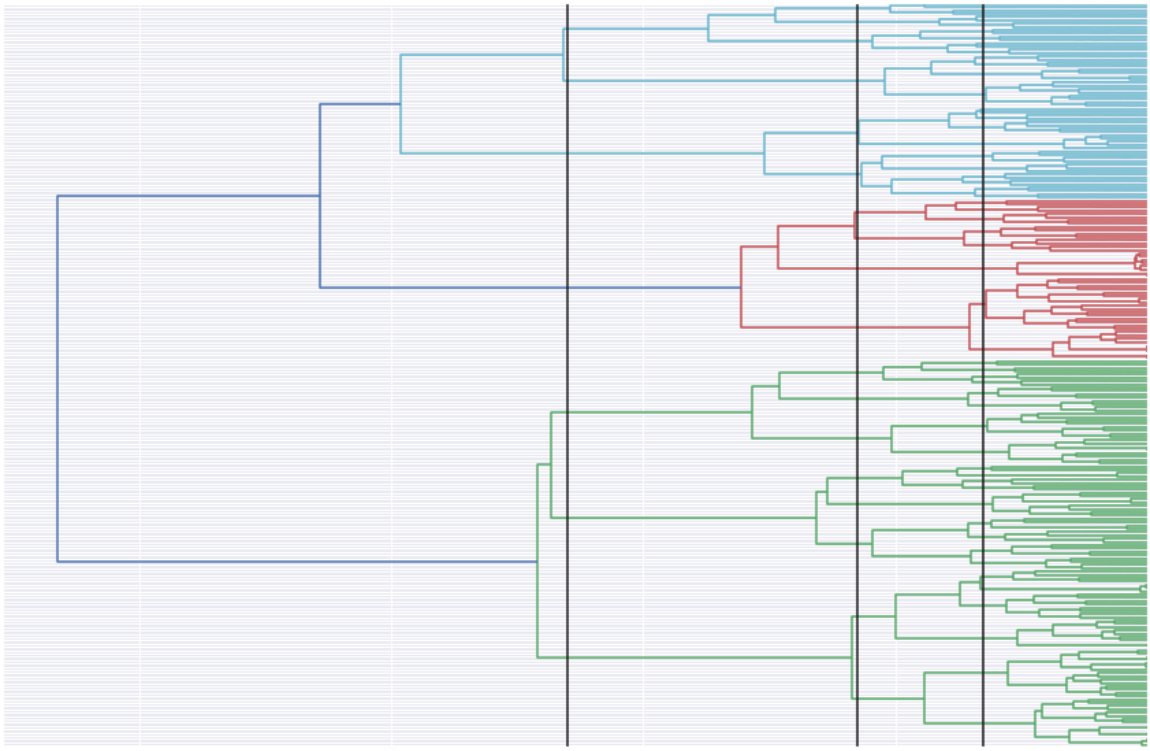


Figure 5.4.: Dendrogram computed on the contact lens dataset. The cutoff values for three detail levels are shown in black.

An example separation of a dataset into large clusters is presented in Figure 5.5.

To make the similarities between patents within a cluster explicit for the user, we summarize key terms from documents within the cluster to a list of cluster key terms. As described in subsubsection 5.2.1.1 “Labels”, patent documents are characterized by a list of the 10 most relevant key terms as computed by TF-IDF. Across the cluster, we count the occurrences of each term within those 10 document terms. The 15 most frequently occurring key terms are considered most relevant for the given cluster. This approach results in more general and common key terms for large clusters, with the specificity growing with each level of detail.

As described in subsubsection 5.2.1.2, we augment cluster key terms with similar words to put them into context and avoid ambiguity. For that, we extract the 10 most similar words from the word2vec model used previously as candidates. Most similar in this case means that the cosine similarities between word embedding vectors are maximal. This is the most computationally expensive step in the pipeline since the similarity has to be computed for every single word in the model’s vocabulary and for every cluster key term. If a candidate appears in more than 10% of documents in a cluster, it is considered an adequate enhancement for the given cluster key term. Since the word2vec model we use only takes single words into account and does not contain embeddings for multi-word units, bigrams cannot be augmented this way.

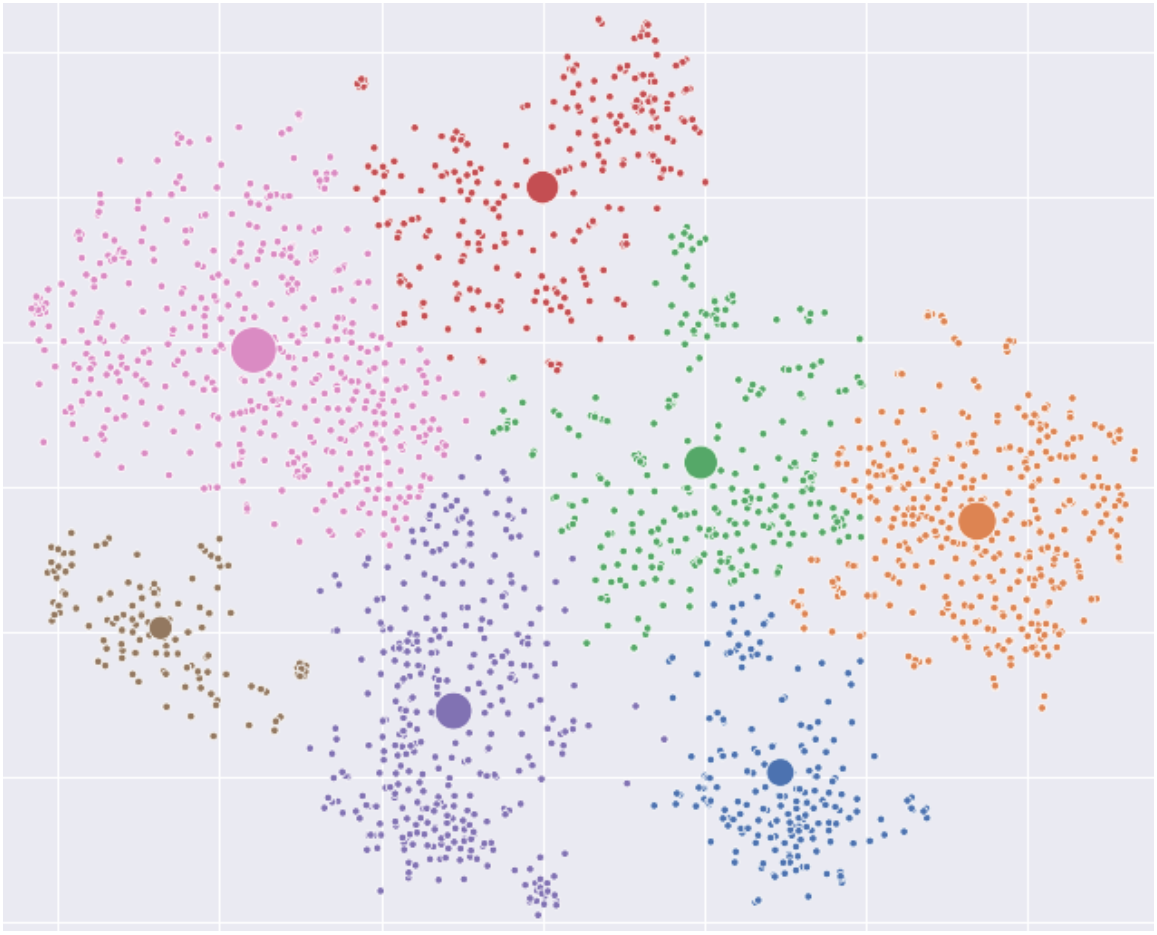


Figure 5.5.: The most abstract detail level (large clusters) of a clustering on the contact lens dataset. Each cluster has its own color. The circles represent cluster centroids and their radius corresponds to the number of documents within the cluster.

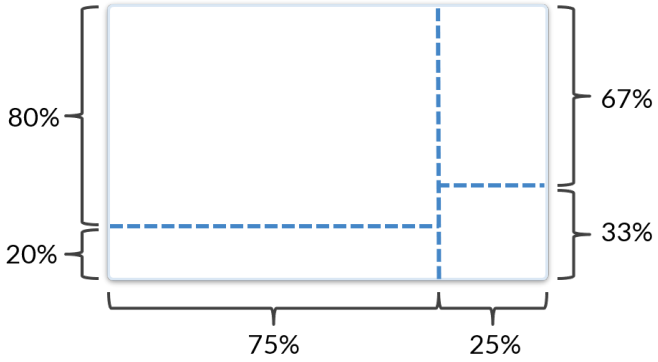


Figure 5.6.: Proportions of rows and columns in the dynamic layout.

The interviews with the patent experts showed that during a patent search, they describe the concept they search for in different levels of abstractness, e. g. umbrella terms and narrower terms (see subsection 3.2.2 and subsection 3.2.4). To aid the user's understanding of key terms we attempted to produce generalizations of cluster key terms which are called *hypernyms*. For example, *chair* is a kind of *furniture*, so *furniture* is a hypernym for *chair*. These kinds of relationships between words in the English language have been manually captured in the WordNet database [4]. Our attempt resulted in very similar hypernyms for all clusters that were too general to be useful, for example *speed, base, length, element, metal, gas, velocity, constant, concentration*. For this reason, we did not pursue this research direction further.

5.2. Implementation of the user interface

In this section, we describe each component of the proposed visualization layout separately. We examine how the dimensions of data are mapped to visual attributes. We then describe how the views are coordinated through the way they react to user interactions.

A schematic representation of the visualization layout is shown in Figure 4.1. The layout consists of two columns and each of them is split into two rows as shown in Figure 5.6. The first column takes 75% of the screen's width and contains the scatter plot (80% of the total height) and the histogram (20% of the total height). The second column fills the remaining 25% of the total width and contains the sunburst with breadcrumbs in the top two-thirds and the detail view in the bottom third. In the early versions of the prototype, the width and height of the layout's elements were fixed. Later, we switched to relative sizes to become independent of the exact screen dimensions. Nevertheless, the prototype is best viewed within a range of resolutions from 1600x900 to 1920x1080 pixels on a screen diagonal from 14 to 24 inches. The main restriction to arbitrary scalability are the font sizes used in the user interface. With resolutions smaller than mentioned above the overlap between text elements is likely to harm readability. With larger resolutions text and point elements will be too small, but it can be alleviated by magnifying the whole web page.

5.2.1. Scatter plot

The scatter plot is the main area of the visualization and is complemented by all other elements: histogram, sunburst + breadcrumbs and detail view. It represents each patent as a point with coordinates that correspond to its position after dimension reduction with t-SNE. Additionally, points are grouped into clusters, each of which is indicated by its key terms. In this subsection we discuss the depiction of single patents first. We then proceed to describe cluster representations.

5.2.1.1. Points

Each data point possesses multiple visual dimensions: position, size, color. In the following paragraphs we describe how they are mapped to dimensions of the data. As mentioned before, the position corresponds to the coordinates in the semantic document space reduced to two dimensions. Size and color are also utilized (see "Size", "Color and glyphs"). They are complemented by connections between patents (see "Connections"). Lastly, each point is labeled with the top key terms of the corresponding patent.

Size

The size of a point depends on the number of forward and backward citations the patent has, all summed up. The radius of the circle varies between 3 and 9 pixels when no zoom is applied. The exact scale used in this mapping is dynamic and dependent on the dataset. The minimum size always corresponds to the lowest number of citations found per patent in the dataset and the maximum size to the highest number. The interpolation between the two values is linear. It is not uncommon for a patent to list hundreds of citations. With this relative scale, we make sure that the size difference is always obvious to the user, regardless of whether the maximum number of citations per patent in the current dataset is 20 or 900.

Color and glyphs

Color of the points is a dynamic variable which is defined by the current state of the sunburst's hierarchy. Specifically, whenever the user navigates to a different sunburst node, colors are newly assigned for its child nodes. Patents that are outside of the scope of the current node are then completely hidden. The remaining points in the scatter plot obtain their color depending on their value of the corresponding metadata attribute.

Earlier iterations of the prototype did not include a solution for displaying patents with multiple values of the given attribute. Instead, they were assigned to the color of the least frequent attribute value. The intuition behind this solution was to provide visibility to a group that otherwise would be less noticeable, especially if those least frequent values only occur in combination with others.

Eventually, we implemented *glyphs* as a solution for the issue of multiple attribute values. "A glyph is a graphical object designed to convey multiple data values" [108]. Usually,

glyphs possess multiple visual attributes such as color, position or length, which are mapped to different dimensions of the data. In our case, however, they are restricted to depict only one dimension of the data, which can acquire one or more values.

Our proposed glyphs are depicted in the form of pie charts with a number of slices corresponding to the number of attribute values. The slices are equally sized since all values of a given attribute are equally meaningful. Since we wanted to keep the patent representations uniformly shaped, a circular multi-colored pie chart was an obvious enhancement of a single-colored circle. The idea was also partially inspired by [34], where pie glyphs show the distribution of topics within a document.

Naturally, the question about legibility of pie charts arose. If they were to contain too many slices, they would be impossible to decipher. To check this, we computed the distribution of how many different values patents included for assignees and IPC classes (see Figure 5.7). The values presented were computed on the contact lens dataset which is described in subsection 5.1.1, but they do not vary greatly between datasets.

On average, there are 1.56 assignees per patent, with an overwhelming majority of patents having only one assignee. IPC classes on the first glance look unsuitable for a pie chart with an average of 4.92 IPC classes per document and a non-negligible amount of patents with more than 20 IPC classes. However, these numbers refer to unique IPC classes throughout the whole IPC hierarchy. In reality, only one IPC level is visible at one time, so we examined distributions after the first subdivision and before the last one to see how many classes truly have to be shown simultaneously. On the subdivision level with single IPC letters (e.g. A or B) there is an average of 1.87 values per patent, and on the group level (e.g. A21B1 vs. A21C3) it is 3.34, respectively. This shows that in total, there is relatively little branching throughout the IPC hierarchy with most of it happening on the last level. This means that the number of slices in a pie chart on each specific IPC level is not too high for intelligibility.

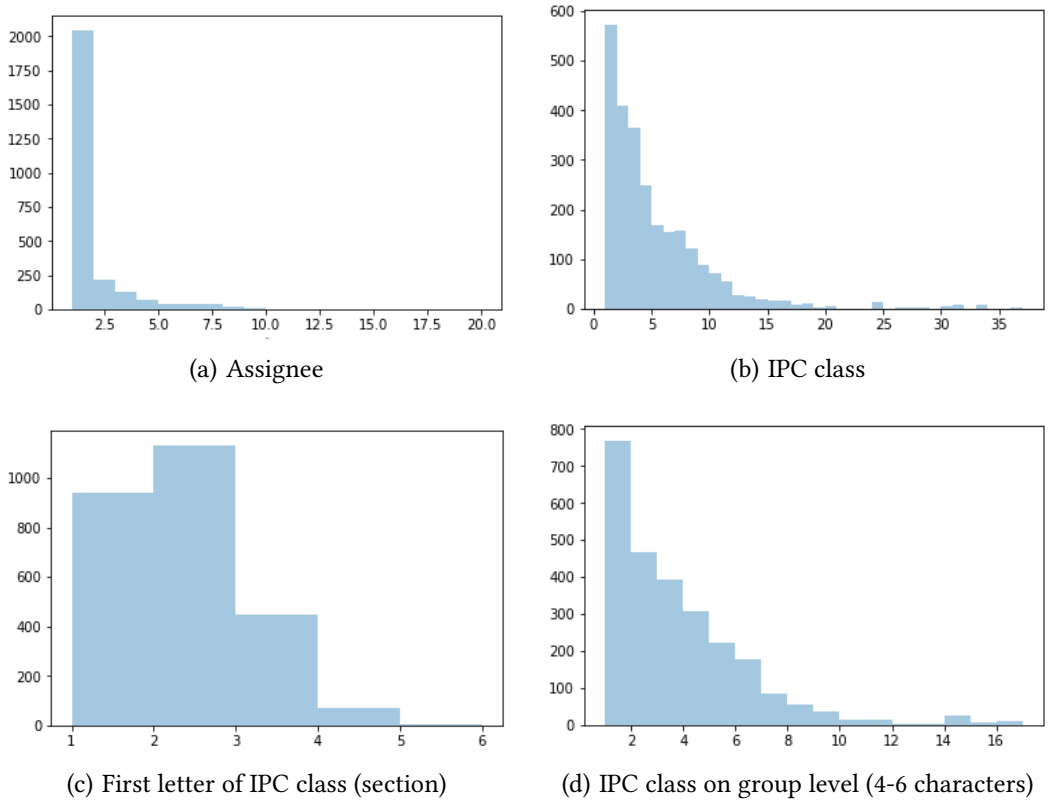
Using glyphs results in continuous areas with the same glyph appearance (see Figure 5.8), which allows the users to make generalized assumptions about the content of those areas. We evaluate how well glyphs fulfill their purpose in subsubsection 6.2.1.2. For simplicity, we refer to glyphs as points whenever the multiple attribute values are not essential to the current discussion.

Connections

Three kinds of possible connections between any pair of patents exist (see Figure 5.9).

As described in subsubsection 2.1.3.2, patent families describe the same invention. Being the same family is the strongest indication of a semantic similarity, which is why families mostly are represented by close-knit groups of points in our semantic approach. We show connections to other family members with black solid lines.

Citations are another kind of possible connection between patents. As opposed to family connections, they do have a direction. Conventionally directed connections are shown with arrows, which is inapplicable in our case. A single patent might have tens to hundreds of citations, which would result in a very cluttered representation if arrows were used.



(a) Assignee

(b) IPC class

(c) First letter of IPC class (section)

(d) IPC class on group level (4-6 characters)

Figure 5.7.: The distribution of the number of values per patent for assignee and IPC class.

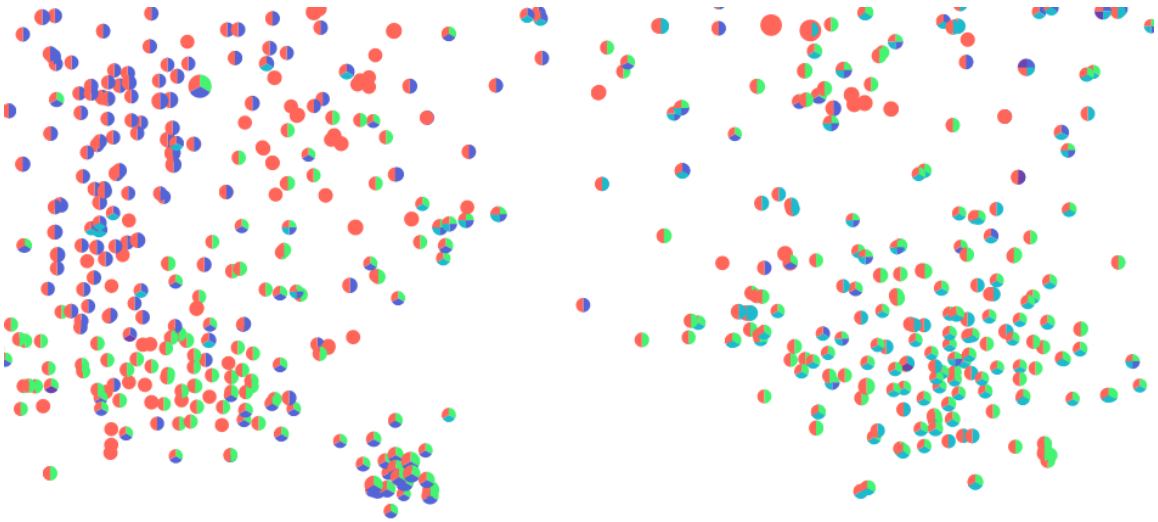


Figure 5.8.: Areas consisting of same kinds of glyphs on the contact lens dataset.

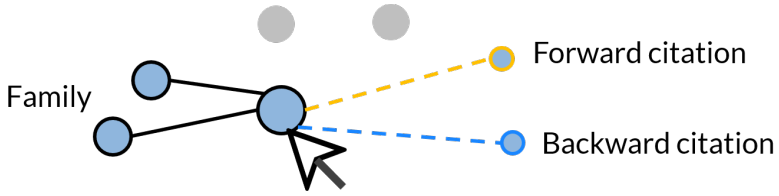


Figure 5.9.: Various kinds of connections between patents.

Moreover, it is useful to be able to see at a glance the areas where the majority of citations come from and go to. As described subsubsection 2.1.3.3, if a new patent application cites an existing patent, it means the inventors are aware of the prior invention and see the novelty in their invention with regards to the prior invention. This is called *forward citation*, which we show with a yellow dashed line. The opposite situation, i. e. from the point of view of an older patent, is a *backward citation* shown with a blue dashed line.

We chose complementary colors (blue and yellow) because they signify exactly opposite things - opposite directions of citation. Yellow is more of an "active" color, which signifies that currently selected patent explicitly mentions the citation. Blue has a more "passive" role, which in our case corresponds to the fact that backward citations are not directly contained in the data, but are instead computed by reversing the connections. Notably, yellow and blue both can be easily seen on black background. It so happens that some patents are listed as both family members and citations, so dashed lines are designed to overlay the black lines and still be clearly distinguishable. Additionally, a dashed line is usually perceived as less important than a solid line, which correctly represents the domain knowledge in this case.

Connections appear while the user is hovering over a patent with a mouse. The user can also choose to select a patent by clicking on it. In this case, the connections persist until the user switches their selection to another patent or resets the selection completely by clicking on the background area of the scatter plot. The selection mode allows the user to highlight a patent of choice and examine its citations and family by hovering the mouse over them. This interaction implements the principles of *focus plus context* (explained in subsubsection 2.1.1.2) and *details on demand* (explained in subsubsection 2.1.1.1). The chosen patent is in focus and its citations and family provide a more detailed representation and also show some context, i. e. what prior art the patent refers to.

Labels

Each patent has a corresponding label that shows up to three top key terms as extracted by the TF-IDF algorithm (see subsection 5.1.5 for details). Information density is an essential characteristic of any user interface that directly impacts usability. To avoid cluttering the visualization space, we use use a heuristic to determine exactly what labels are shown and how many top key terms they include.

Our experiments showed that about 250 labels (consisting of one key term) for points can be shown simultaneously and remain mostly readable. So we decided to limit the labels to a maximum of 250. This means that some points are shown unlabeled until the user

restricted the area they are interested in to under 250 patents. After each operation, such as filtering, panning or zooming, we count the points that are currently situated within the viewport. We then divide that amount by 250. If the resulting quotient q is over 1, we round it up to the next integer to get n. In this case, every nth point is labeled with its top key term. For example, if 980 patents are currently visible, every 4th of them gets labeled, and three-fourths of patents are shown with just a point without a label. If q is between 0.7 and 1, every patent in view is labeled with its top key term. Usually, this happens during the examination of small clusters (see subsection 5.1.7 for explanation of three cluster sizes), when the user's attention shifts to a single document. For values of q between 0.3 and 0.7, there is sufficient space for top two key terms and for values under 0.3 for three top key terms for every patent.

The above-mentioned value intervals are chosen to maintain a visual balance between points and their labels and to minimize overlapping text. If the user wishes to examine further key terms beside the top three ones, they have the possibility to inspect them in the detail view (described in subsection 5.2.4) along with complete information about the patent. We provide the possibility for the user to comprehend the distribution of points and their colors without distraction before starting with the detailed analysis. To support this, we make all point labels invisible when the zoom level is less than 1.15.

The varying number and length of patent key terms result in a dynamic level of detail. It is one of our multiple embodiments of the *semantic zooming* mechanism (explained in subsubsection 2.1.1.4). The evaluation (see subsubsection 6.2.1.5) showed that our heuristic resulted in a readable representation for multiple levels of detail.

Zooming

Zooming causes a multiplicator to be applied to the point size. The multiplicator value varies from 1x to 1.7x and is interpolated linearly depending on the exact zoom level. The maximal possible zoom level is 10, but after it reaches 3, points and text in the scatter plot stop increasing in size, so further magnification only increases the distance between the points. Thus, we intentionally increase the amount of white space to allow the user to focus their attention on specific patents. Also, at this detail level, the documents are accompanied by a list of key terms which need to be readable. Improved readability is also a reason for the additional white space.

The lines representing families and citations also change subtly with the zoom level. Their width changes from 2 to 3 pixels to stay in proportion with the point size.

To allow users to quickly go back to the overview of the dataset, we added a "Reset zoom" button in the latest iteration of the concept. This way, the users are able to go set the zoom level back to 1 with one click instead of turning the mouse wheel two to seven times depending on the size of the dataset. We would like to mention that navigating to the maximum zoom level is rarely, if ever, continuous. The user pauses to analyze the currently presented information to steer their further examination. Therefore zooming into the dataset is not as cumbersome as zooming out without using the reset button would be.

5.2.1.2. Clusters

As described in subsection 5.1.7, we cluster patents hierarchically based on their distance in the 2D space, i. e. their proximity in the scatter plot. From the computed agglomerative clustering we pick three specific cluster configurations, which provides us with three levels of detail. We refer to them as *large, medium* and *small* cluster sizes.

Clusters are represented in the visualization by their key terms. Since we are essentially generating a themescape, we considered a solution from the domain of map drawing. On a map, labels for areas such as forests, deserts or lakes are often not straight but bent to match the shape of the area. We tried to replicate this behavior by approximating the points in the cluster by a polynomial curve. The cluster text was supposed to stretch and follow the curve to make the shape of the cluster visible. The result of our attempts is presented in Figure 5.10.

At that stage, we were computing the whole clustering based on distances in the high-dimensional document space. With this approach, dimension reduction to 2D resulted in elongated cluster shapes which, as we initially hoped, would be easy to approximate. At the same time, elongated clusters were not separated clearly and often overlapped, which significantly reduced the quality of the approximating curves and the readability of text placed on them. It was also difficult to adjust how the text should stretch depending on the length of the key terms and the length of the curve. For these reasons, we decided not to pursue this direction of research further. Notably, [98] uses the professional Geographic Information System (GIS) software ArcGIS extended by the Maplex labelling engine to place labels on their themescapes with great success (see Figure 2.2 for an example themescape). This speaks strongly for the potential of applying map drawing methods for non-geographical data.

In the current embodiment of our approach, each cluster is represented by its top three key terms situated exactly in the middle of the cluster. To place them in a compact way, we show the most relevant term in the middle and complement it by the second and third terms above and below it. The font size of the top term is bigger by three pixels to emphasize its importance according to principles of visual hierarchy. Moreover, both visibility and font size differ between large, medium and small cluster labels depending on the current zoom level (see Table 5.1).

Cluster level	Font size, pixels	Visible at zoom level
large	23 to 28	<1.6
medium	18 to 22	1.5 to 2.1
small	14 to 18	2.0 to 3

Table 5.1.: Font sizes for different cluster sizes

As the user zooms into the scatter plot, bigger clusters are replaced by smaller ones. This continues until the zoom level of 3, when the cluster labels disappear completely allowing the user to examine specific patents efficiently. Notably, the visibility intervals overlap

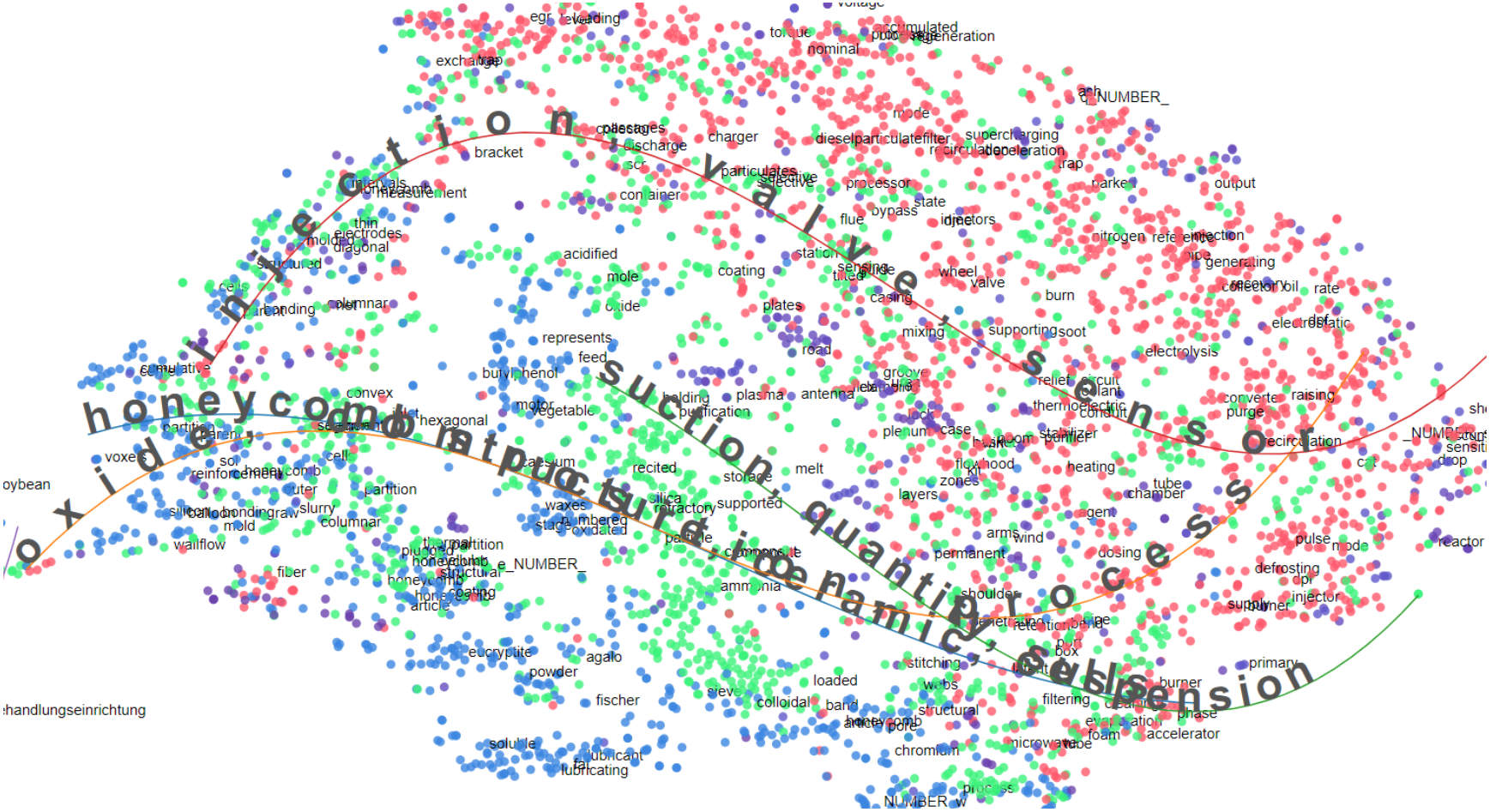
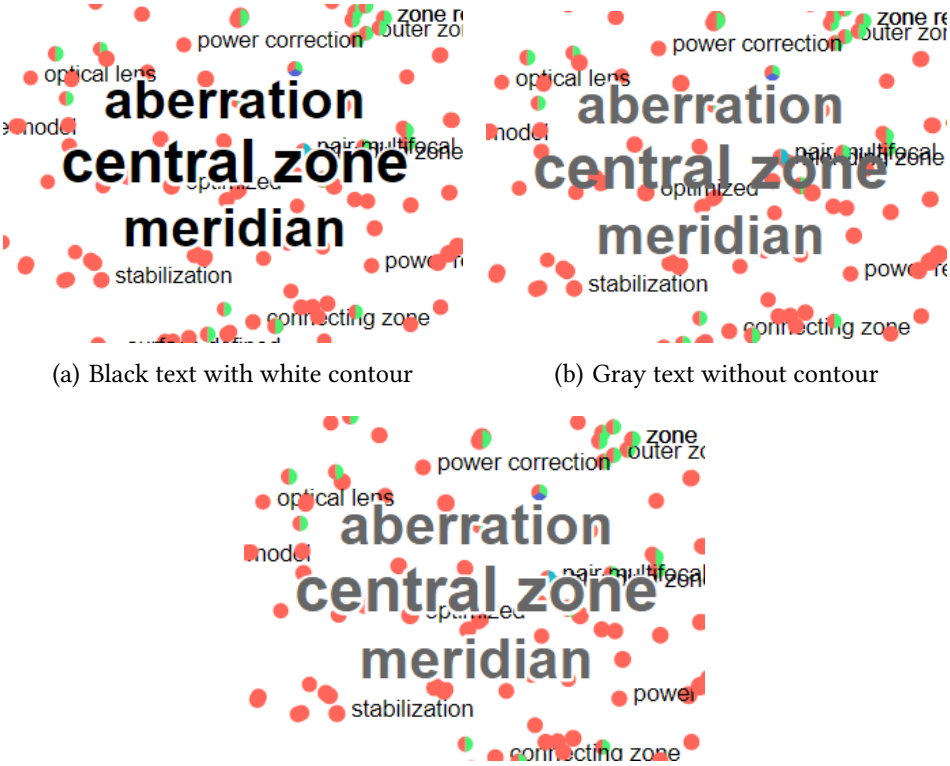


Figure 5.10.: Cluster shapes approximated by curves on the diesel engine dataset. The text for cluster labels is placed on the curves.



(a) Black text with white contour

(b) Gray text without contour

(c) Gray text with white contour

Figure 5.11.: Font color and contour increase readability of cluster labels.

slightly. This means that briefly, both large and medium or both medium and small cluster labels are seen. The idea here is to make the transition between different levels of detail smoother. The font size of cluster labels increases slightly while zooming in. As mentioned before, points themselves and connections between points also increase in size. This behavior is implemented in a consistent way throughout the scatter plot to support the feeling of "moving into" the dataset.

To optically balance out the increasing line thickness as the clusters labels become larger, we vary the font color slightly. Labels for single patents are black, so we made little cluster labels a few shades lighter, which yields a dark gray. By analogy, medium and large clusters were also assigned gray color slightly lighter than at the corresponding previous level. A comparison of black and our lighter alternative for large cluster labels can be seen on figures 5.11(a) and 5.11(c). Additionally, the distinction between cluster terms and patent terms is accentuated by using unequal colors. To strengthen the effect and increase readability on a colorful background, we draw a thin white contour around the cluster labels (compare figures 5.11(b) and 5.11(c)).

Some of the cluster key terms are assigned a list of *augmenting words* which provide context for occasionally unclear or ambiguous terms (see subsection 5.1.7 for details on augmentation). This information is not of a high-priority, thus we only display it on demand. A tooltip with a list of augmenting words appears when and only when the user hovers over the corresponding term with the mouse. No tooltip is shown when the

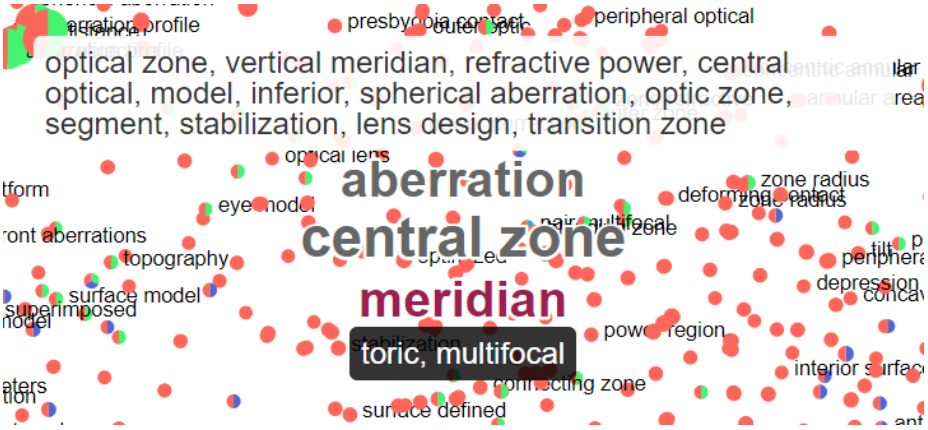


Figure 5.12.: Tooltip with augmenting words below and tooltip with all cluster terms above

given term possesses no augmenting words. Naturally, there has to be a clear indication of what term out of three possible presented is enhanced with a tooltip. There was no easy technical possibility to show the tooltip on the side of the corresponding term, so we placed it underneath the cluster label. We make the current state of the interface explicit by highlighting in wine red color the term which the tooltip currently corresponds to as shown in Figure 5.12.

Since we actually computed the top 15 key terms per cluster but had only been showing 3 of them so far, we decided to display the remaining 12, too. For that purpose, in the latest iteration of the prototype we added another tooltip, this time above the three terms, with distinctly different appearance to avoid any confusion. The interaction follows the same pattern as with the first tooltip, i. e. it is displayed on mouse hover.

5.2.2. Histogram

The histogram view shows the number of submitted patent applications per year in the form of a bar chart (see Figure 5.13). It is, in fact, a histogram in which the bin size equals one year. It allows the user to involve the temporal dimension into their perception of the data. With a brushing interaction, the user can filter out the data outside the selected interval. The selected window can be moved or expanded with the help of the handles on either side. With a click on the background of the histogram the selection can be reset. On the X-axis, only every second year is labeled to avoid overcrowding. To compensate for that, the years of a current selection are shown in the upper-left corner. It helps eliminate the need for mental computations as the user chooses a time interval of interest.

The values in the histogram are computed based on the current top node in the sunburst. Initially, the top node corresponds to the whole dataset, and then switches to its specific portions as the user focuses their attention on specific metadata attribute values. The scale of the Y-axis of the histogram is adjusted dynamically as its maximum value changes. Additionally, the histogram is directly involved in other forms of interactions between views that are described in subsection 5.2.5.

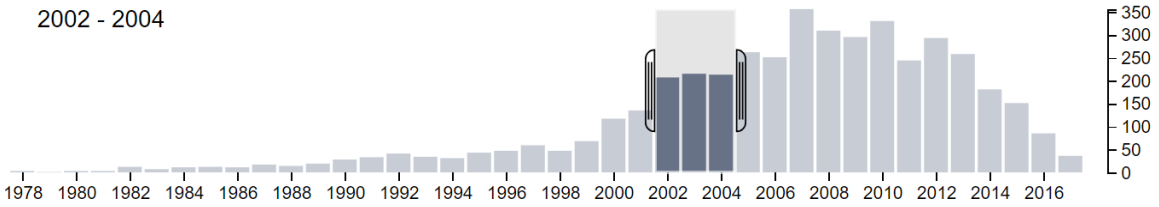


Figure 5.13.: Histogram view as seen on diesel engine dataset. The time interval from 2002 to 2004 is selected.

5.2.3. Sunburst and breadcrumbs

5.2.3.1. Sunburst

Our implementation of sunburst is based on interactive demos from [8] and [106]. The angle taken by nodes is linearly dependent on the number of patents that have the corresponding attribute value. The nodes are sorted by their value in the descending order. When there is enough space available, node names are displayed in the center of a node. The text follows the node's arc and is surrounded by a white contour for a better readability. The title of the node and an absolute number of patents belonging to it are also shown as a tooltip when the user hovers over the node.

The color of nodes on the first sunburst level is based on a cyclical rainbow color palette. Basically, the angles within the interval $[0, 360]$ are mapped to a corresponding position from $[0, 1]$ on the color scale. A node's angle for the purpose of this calculation starts at 0°and ends in the middle of the node's arch where its title is. Because the color scale is cyclical, nodes nearing 360°have similar colors to those near 0°. Child nodes expanding from the first sunburst level are assigned a spectrum of shades from darker than parent to lighter than parent. This emphasizes that they belong to the parent.

When patents possess multiple values of the same attribute, for example, assignee or IPC class, occurrences of each single value are bound to exceed the number of patents when summed up. To represent that overlap correctly, we normalize the values so that they add up to a total of 100%. For example, assume there are two assignees A and B in a dataset, 80% of patents have A as their assignee and 40% have B. This means that 20% of patents have been submitted by A and B together. To correctly represent the relative distribution, we would draw node A as $\frac{80}{80+40=120} = 67\%$ of the sunburst and node B as $\frac{40}{120} = 33\%$.

It is possible to navigate to deeper hierarchy levels by clicking on the chosen node. To go back one level, the user needs to click on the circle in the middle of the sunburst which represents the current top node. The transitions between two states are animated to emphasize the change in the system state: sunburst nodes fold and unfold like a fan. Change in the state of the sunburst also directly affects the scatter plot: first, only points belonging to the currently chosen node stay visible, second, their colors change to match the new sunburst nodes. In addition to clicking, the user can also hover over a sunburst node to see a preview of their choice. In this case, not related patents are hidden in the same way as with clicking, but point colors stay as they are because the sunburst nodes

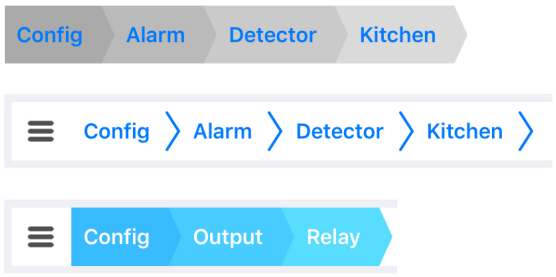


Figure 5.14.: Different kinds of breadcrumb design. Image source:[20]

have not yet changed their colors. Only the nodes on the path to the currently highlighted node retain their color. Remaining nodes are shown with a lighter color to keep all focus on the highlight. The highlight is reset after the user moves the mouse outside of the sunburst's circle.

5.2.3.2. Breadcrumbs

Visibility of system status and *recognition over recall* are two of usability expert Jakob Nielsen's ten heuristics for user interface design [77]. To take them into account, the first change from the initial concept was adding *breadcrumbs* to the sunburst view.

Breadcrumbs are a metaphor most familiar to users from website navigation (see Figure 5.14 for examples). The name originates from the German fairy tale about Hansel and Gretel, who left a trail of bread crumbs in the woods to be able to find their way back [63]. Breadcrumbs are applicable with hierarchically arranged navigation, i. e. when there is only one possible path to every node in the navigation tree. According to [43], they are usually used as an optional aid to navigation and should be less prominent than the primary navigation element (which is the sunburst in our case).

With the first implementation of the sunburst, it soon became clear that some hierarchy nodes were too narrow to include their title within. Moreover, the titles that could be shown were placed at different angles, which prevented a sequence of highlighted nodes from being read easily. Breadcrumbs build a straight line, which increases readability. Additionally, when a user switches to a deeper level of the sunburst hierarchy, the parent nodes are no longer visible. In accordance with the *recognition over recall* principle, it should not be expected of any user to remember what parent nodes came before. We included breadcrumbs as a visual aid that clearly and consistently represents the status of the system and reassures the users of the result of their actions, especially when they interact with small sunburst nodes. The result can be seen in Figure 5.15. A percentage on the right side of the breadcrumb trail represents the fraction of patents from the currently highlighted hierarchy node with regard to the current root node. It is especially useful when there is a significant overlap between node values because it shows the true value of the node independent of its normalized angle.

Short codes for IPC classes fit well into the width of the line allocated to them. To keep the breadcrumbs view concise for assignees as well, we chose only to show the first nine

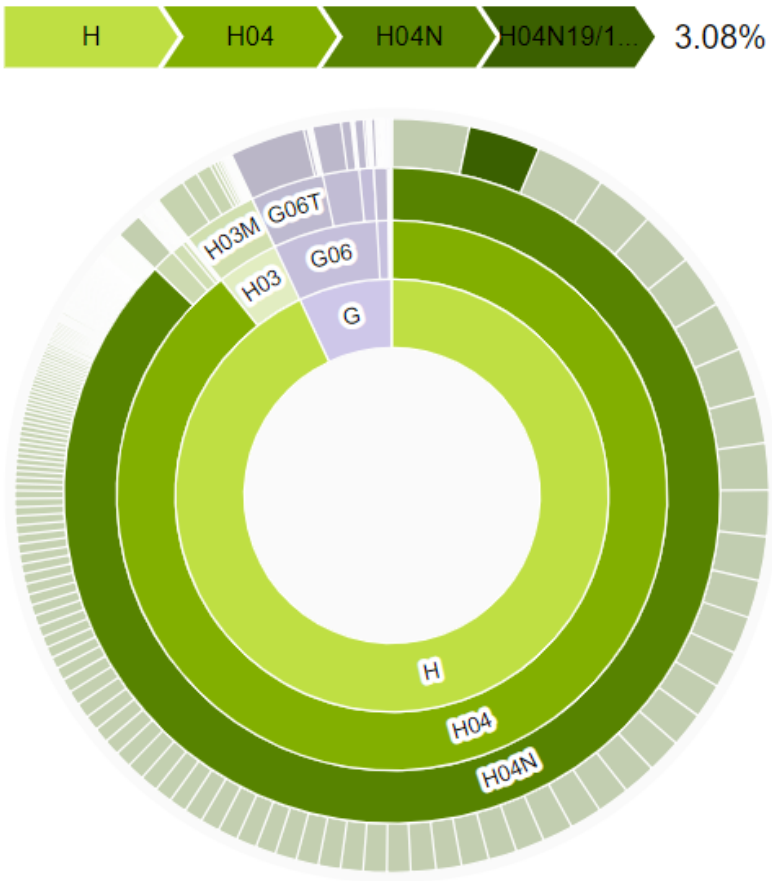


Figure 5.15.: First version of breadcrumbs complementing the sunburst view.

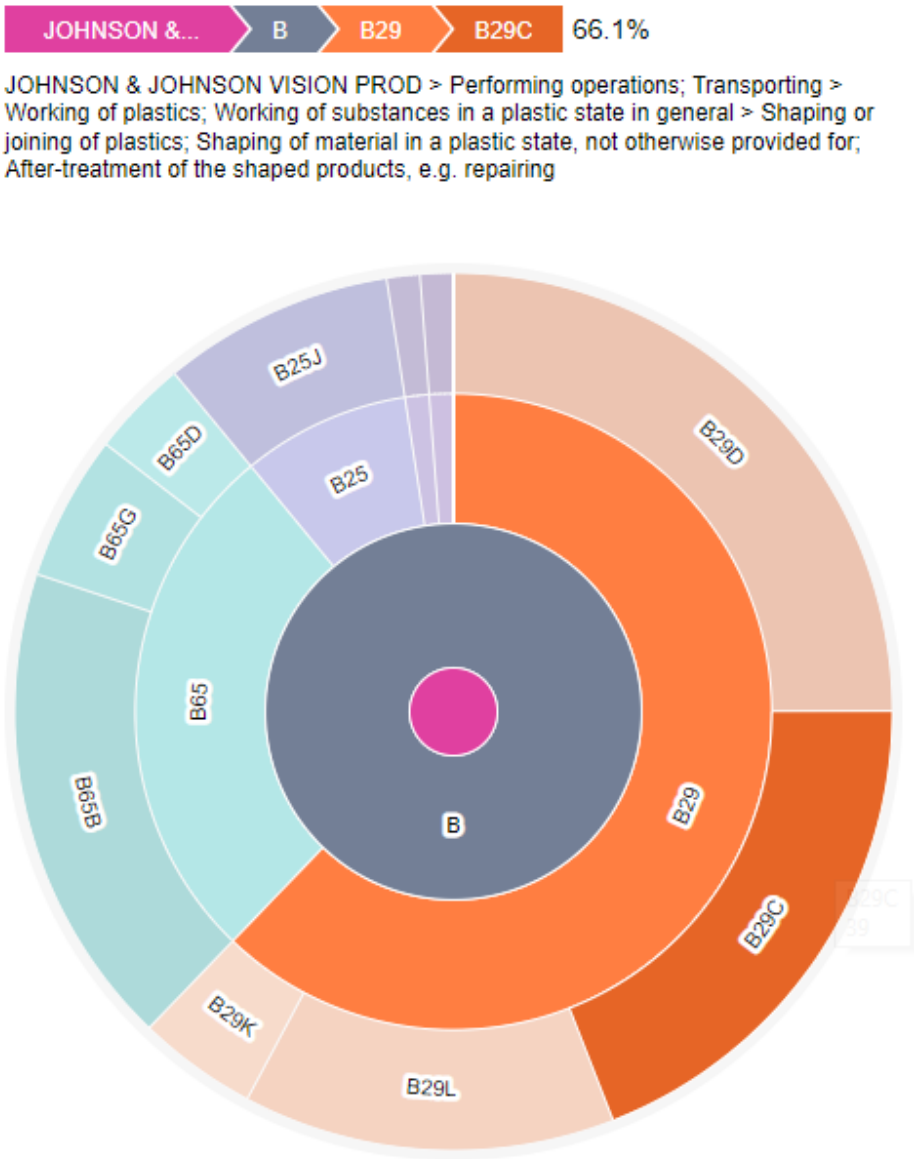


Figure 5.16.: Full-text titles of sunburst nodes added to breadcrumbs.

characters of a node's title followed by an ellipsis mark. As assignee names can be over 30 characters long, the need to see full node titles remains. The idea proposed by patent experts during a feedback meeting helped address this issue. The experts wished to see full descriptions for IPC codes, for example, "optical elements, systems, or apparatus" for class G02B (see [49] for full schema). This led us to complement the graphical brief breadcrumbs with a full-text part as seen in Figure 5.16. Full assignee names could then be shown completely along the IPC descriptions.

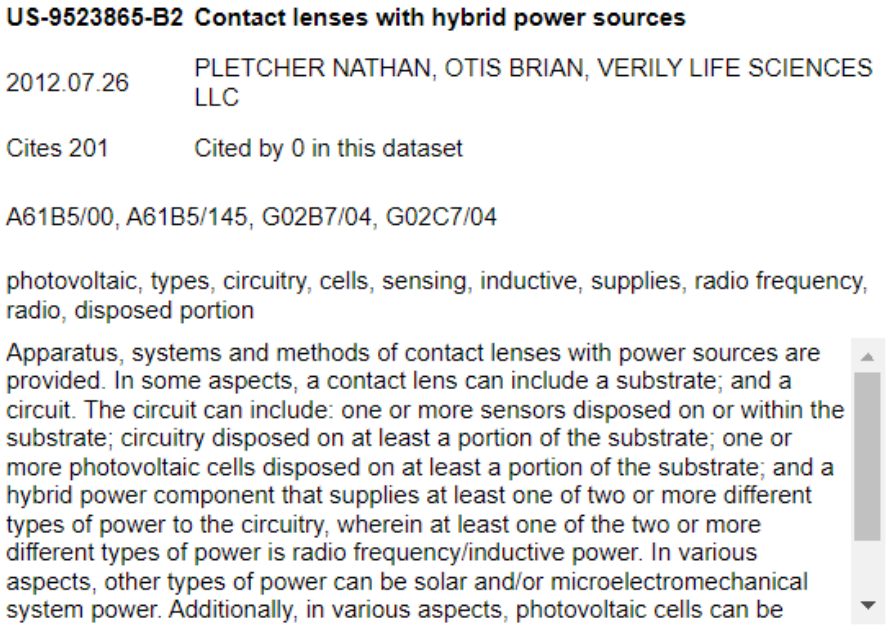


Figure 5.17.: Detail view on an example patent from the contact lens dataset.

5.2.4. Detail view

The detail view follows the principle of *details on demand* (see subsubsection 2.1.1.1 for details). It allows the user to examine all of the available metadata per patent. The information is organized in a tabular manner for compactness (see Figure 5.17). Included are (left to right, top to bottom) publication number, title, application date, a list of assignees, forward and backward citations, a list of IPC classes, a list of the top 15 relevant key terms and the abstract.

If the user would like to study the patent text thoroughly, they can double-click anywhere in the detail view. This causes an additional browser window to appear, which contains the full text of the textual parts of the patent, i. e. abstract and claims (see Figure 5.18). When the user double-clicks on another patent, the already opened window persists and a new one is opened. This permits a detailed examination and comparison of multiple patents.

5.2.5. Interactions between views

For a consistent behavior across all parts of the interface, we enable similar kinds of interactions for the sunburst, the histogram and the scatter plot. We distinguish between three kinds of interactions:

- hovering over an object for a preview of the changes - “Highlighting”
- clicking on an object to make the change persistent - “Selecting”

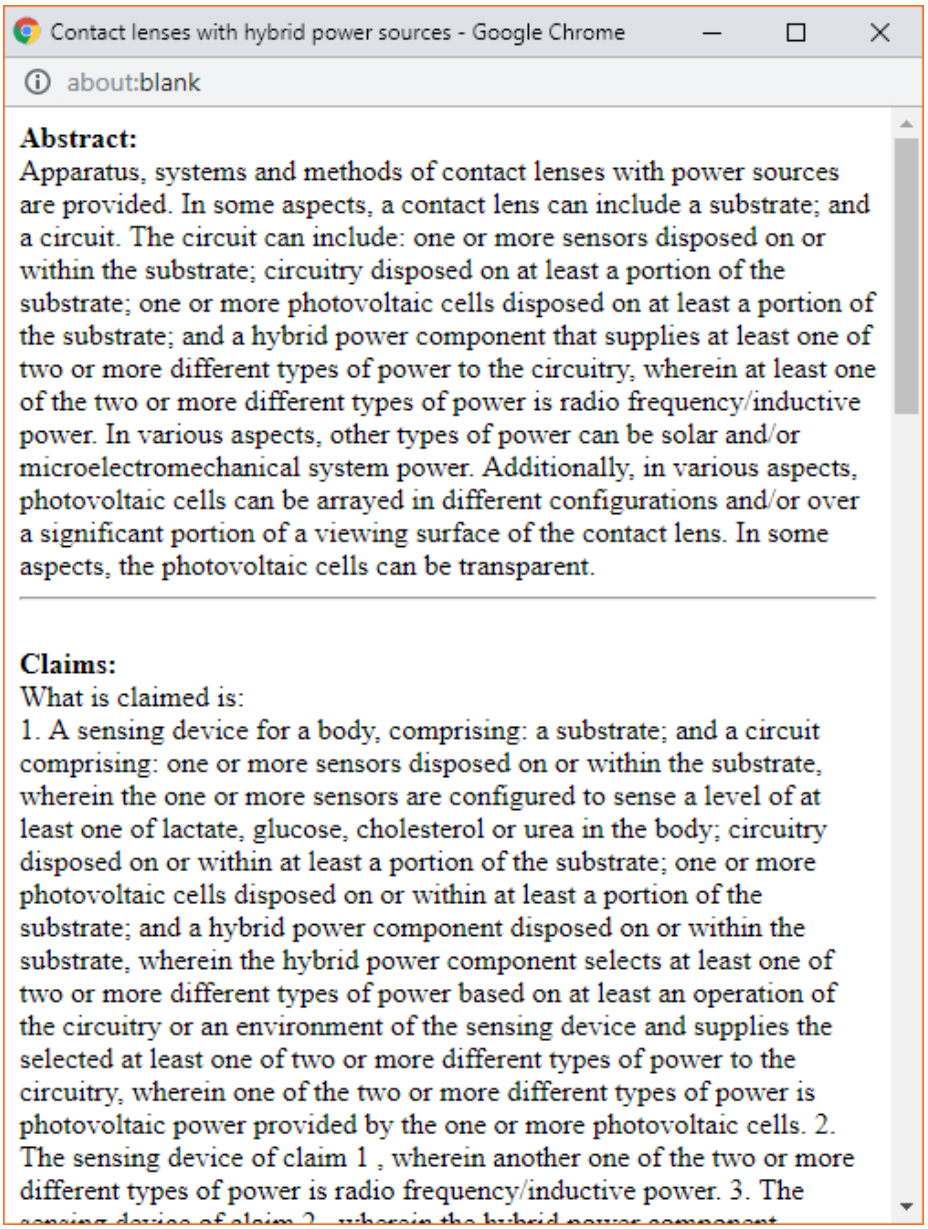


Figure 5.18.: Window with full text (abstract and claims) of an example patent from the contact lens dataset.

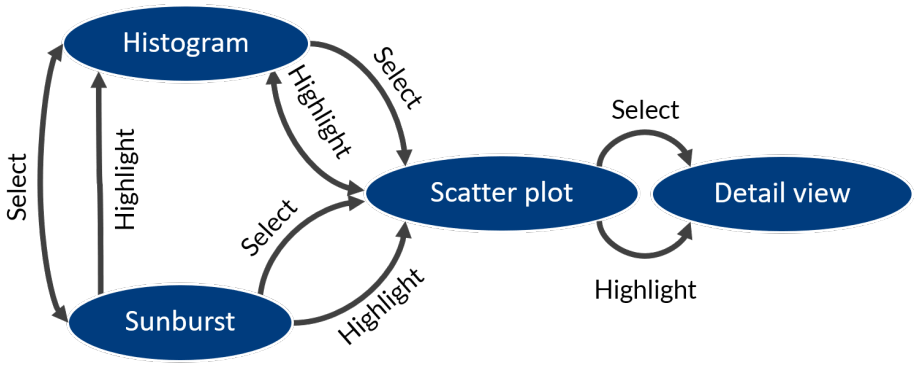


Figure 5.19.: Diagram of interactions between views. The arrows point from the view where the given interaction happens to the view where it takes effect.

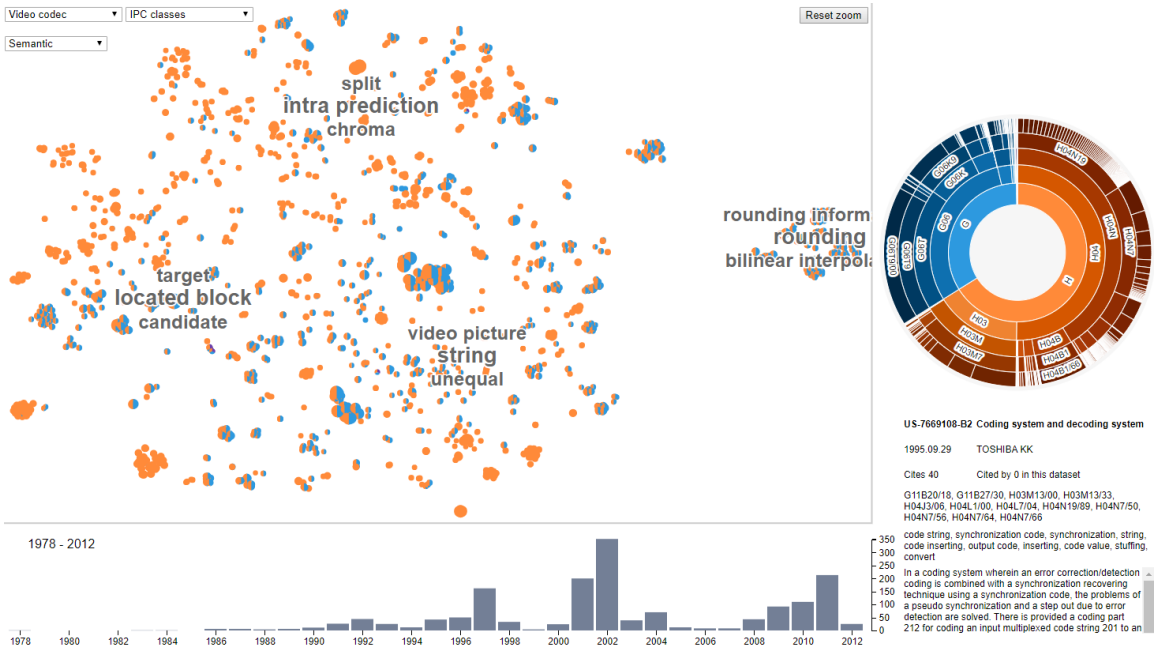
- clicking on a background of the view to reset the selection.

An overview of how those interactions influence coordinated views is shown in Figure 5.19. In this section we describe the possible user actions that had not yet been covered in the previous sectionss.

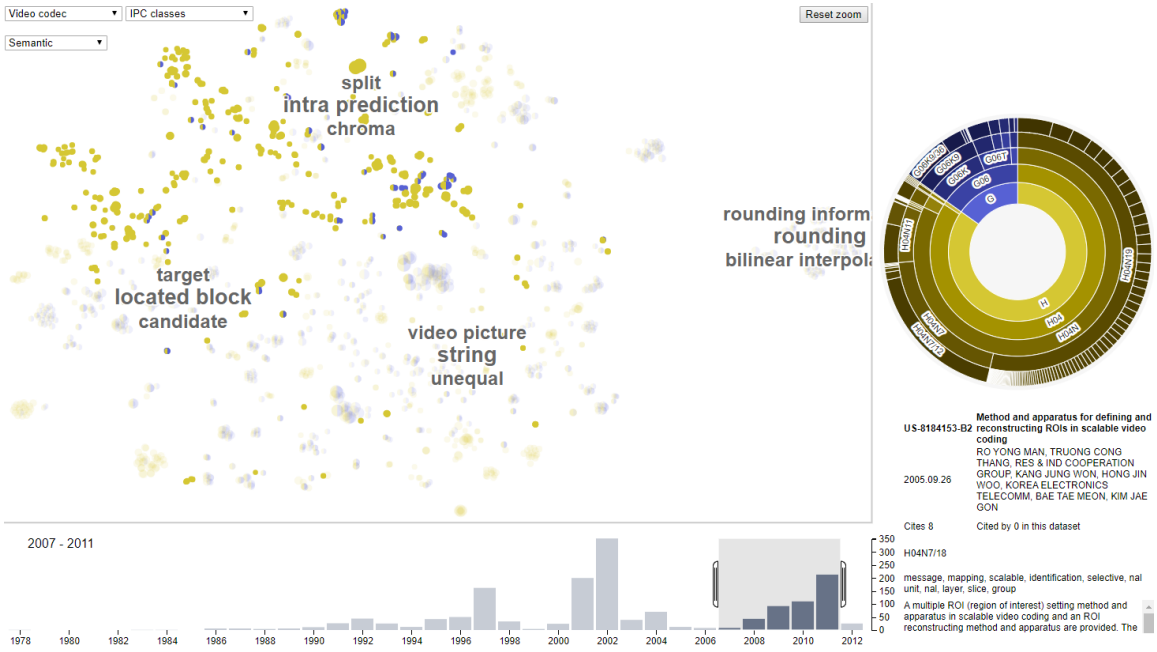
On a histogram, the user can select a certain time interval by brushing. The histogram bars for the years outside the selection become grayed out, so do the patents submitted outside the selected interval. Then, the sunburst is generated anew based only on the patents within the selection. The colors of the remaining points then adjust to match the new state of the sunburst. See Figure 5.20 for a comparison of states before and after brushing. In this example, one can see that patents submitted from 2007 to 2011 are concentrated in one thematic area. Moreover, less of them are assigned the IPC class G - “Physics”, while H - “Electricity” becomes more widespread. This might be an indication of a change in the meaning of the IPC classes over time: the first version of IPC classification was developed in 1968 before the rapid development of information technology. The selection in the histogram can be reset by clicking on its background.

When the user hovers over a sunburst node, this node’s contribution to the histogram is displayed in the color corresponding to the sunburst node. In other words, the histogram becomes a stacked bar chart in which the bottom part of the stack corresponds to the highlighted sunburst node and the upper part of the stack includes all other patents. This interaction allows the user to follow the temporal trends in the development of a single IPC class, assignee or country. For example, Figure 5.21(a) shows that 3D printing technology has started developing rapidly in China since 2015. If the user then clicks on China, it moves to the center of the sunburst and its children take over the whole circle. The bars corresponding to China that were blue on hover now constitute the whole histogram and the Y-axis is scaled accordingly (see Figure 5.21(b)). Moreover, the points in the scatter plot now match their colors to the child nodes of China.

The scatter plot also implements the highlighting and selection with regard to single patents. As long as the user hovers over a certain point, it is emphasized by a black contour and its citations and family members become visible. The detail view then shows the metadata of this particular point, but the information persists even when the mouse is

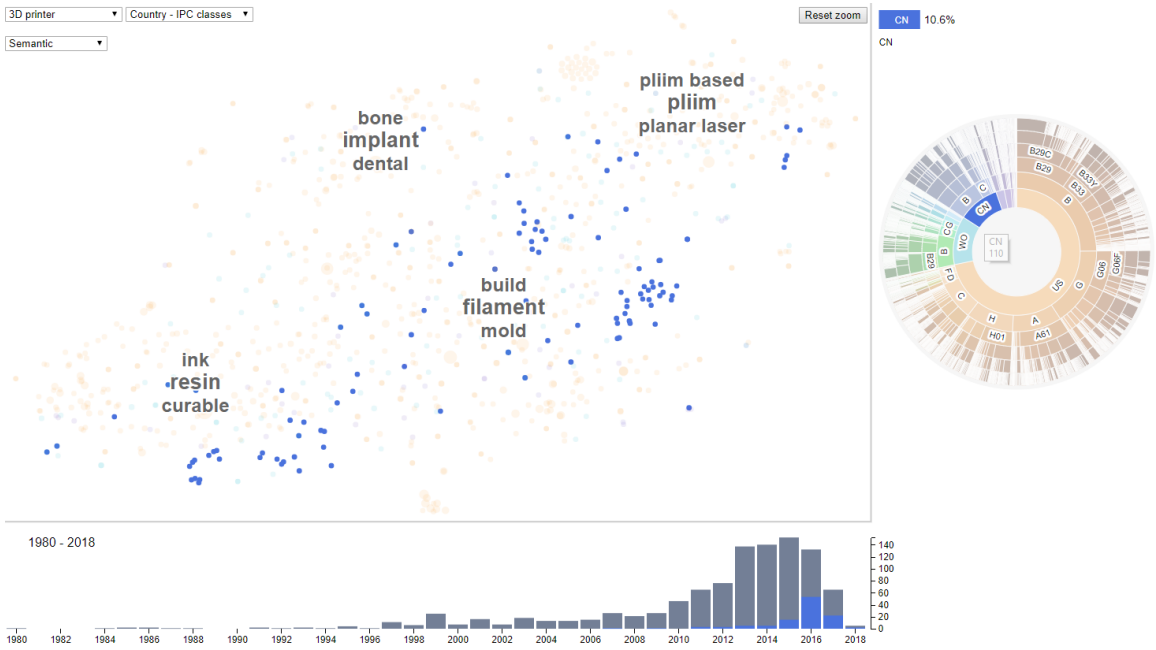


(a) Before brushing

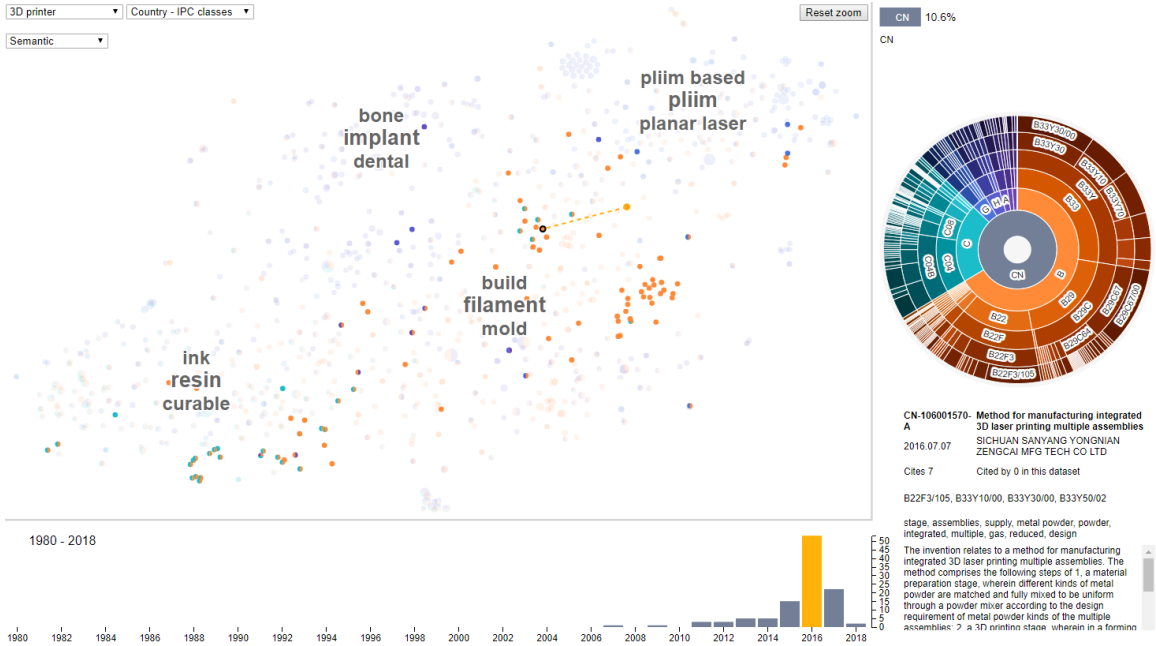


(b) After brushing

Figure 5.20.: The impact of a brushing action on a histogram on the scatter plot and sunburst. Video codec dataset



(a) While hovering



(b) After clicking

Figure 5.21.: The impact of hovering and clicking on a sunburst node on histogram and scatter plot. 3D printer dataset

taken away until some other patent is highlighted. Moreover, the year when the patent was submitted is accented in bright yellow on the histogram (see Figure 5.21(b)). It gives the user a quick impression about the age of the patent without the need to read this information in textual form in the detail view. Clicking on a patent makes its connections permanently visible. The user is then able to examine related patents by hovering over them. The detail view is reset to the details of the selected point when no other point is hovered over. A click on a background of the scatter plot resets its selected point.

All above-mentioned interactions between linked views taken together allow the user to focus their attention on any desired aspects of the data, be it the temporal aspect or specific metadata values. The display can be restricted to a region of interest by various filters. We made an effort to enable interactions on both micro- and macrolevel (for single patents and for groups of various sizes). Users' understanding of the interplay between views is evaluated in subsubsection 6.2.1.4.

6. Evaluation

The objective of our work is to provide a solution for the problem of exploration of large document collections. To examine how well our proposed visualization meets the goal, an empirical evaluation is necessary. This we achieve by conducting a summative user study, which is a part of the case study described in chapter 3. The purpose of this stage and the procedures for data collection have been decided upon as described in subsection 3.1.2. In this chapter we describe the execution of the summative study and discuss its results.

6.1. Procedure

Four patent experts agreed to take part in the summative study. Three of them had already participated in the formative study described in section 3.2.

Being asked to perform tasks while being recorded might be stressful for the participants and therefore might influence the results negatively. It is therefore recommended to preserve the context of the participants' usual workday routine as much as possible [82]. For this reason, the study took place in a meeting room at the participants' workplace. The participants were invited to solve tasks and talk about their experience while using the prototype in a one-on-one conversation.

Each individual appointment started with an explanation of the plan of the study: first a short introduction, then the first part focused on evaluating the usability of the prototype as a whole, followed by a second part focused on comparing the proposed approach with the baseline. The participants were encouraged to talk while they solved the tasks to describe what they were doing. The full plan of the study can be found in section A.4.

As a preparation for the actual study, two test rounds were conducted with volunteers with backgrounds outside the patent domain. This experience allowed us to experimentally detect inefficiencies and inconsistencies in the study plan and improve it before the proper study began. The examiner had to multitask during the study:

- acknowledge the participants' statements and encourage them to think aloud
- make sure that not too much time is spent on one task
- look out for technical problems
- guide the participants through the study procedure

For this reason, the practice rounds were immensely helpful to rehearse following the procedure of study plan exactly.

All four experts familiarized themselves with the ideas behind the visualization approach and with the prototype itself. This took place during the feedback meeting where the first proof-of-concept prototype was presented and also during the mid-term presentation for the thesis. It allowed us to keep the introduction part short and to establish the rapport with the user quickly.

During the introduction, possible interactions with the prototype were explained and users were encouraged to try them out themselves immediately. For that part, a dataset about 3D printers was used. This way, only the knowledge about how to control the prototype was transferable to the following tasks.

After the first task-solving part, a SUS questionnaire was offered. Some participants voluntarily commented on their answers. Then a second task-solving part began. With the study being in-subject, every participant evaluated both the semantic and the baseline approaches. To counter learning effects, half of the participants tested the baseline approach first, then the semantic approach. The other half started with the semantic approach first and finished with the baseline approach.

After the subject finished the tasks pertaining to one approach, a questionnaire was offered to capture impressions about that approach. The participant was encouraged to speak about why they answered in the way they did. The study ended with a brief general discussion of the participants' impressions, any perceived advantages and disadvantages of the prototype. All relevant guidelines for the interviews already mentioned in subsection 3.2.1 were followed: intentional pauses to encourage participants to speak, open questions instead of leading ones, etc.

Starting after the introduction, the study was recorded with the help of a screen capturing software OBS Studio. The participants' actions and their commentary were captured and later transcribed for the evaluation. The resulting script is 23 pages long and is not included in the thesis for brevity's sake and for the protection of the participants' privacy. The transcription enabled us to make conclusions regarding the hypotheses we describe in subsection 6.2.1.

6.2. Results

6.2.1. Think-aloud

We used the contact lens dataset described in subsection 5.1.1 for the evaluation. This domain was chosen partly because of the participants' background in chemistry and medicine. Nevertheless, it is important to remember that the participants were not closely familiar with the subject of contact lenses.

We put forth a number of hypotheses to evaluate different aspects of the proposed approach. The hypotheses from subsubsection 6.2.1.1 to subsubsection 6.2.1.5 pertain to usability of the prototype. Evaluating them shows whether there is a match between the user interface and users' mental model of the patent landscaping task. The last two hypotheses

in subsubsection 6.2.1.6 and subsubsection 6.2.1.7 have the specific goal of comparing the baseline approach with the semantic approach.

6.2.1.1. Hypothesis 1: Color mapping of points depending on sunburst is understood and helps identify clusters

Color is a pre-attentive visual attribute, which means differences in color are processed effortlessly and in parallel without any attention being focused on the display [104]. Therefore, we expect that mapping categorical attributes, such as IPC class, to colors helps experts identify coherent areas.

There were two tasks for evaluating this hypothesis:

- Task 1. What IPC classes (on the section level, one letter) appear together often?
- Task 11. Briefly describe broad thematic areas in the dataset (big clusters). Evaluate the positions of the clusters relative to each other.

During Task 1 the participants implicitly grouped same-colored points

Co-occurring IPC classes yielded continuous areas with the same color combinations for both the semantic and the baseline approaches. Assuming that the IPC classes were assigned consistently, this indicates that both approaches were able to capture thematic groups within the domain. All four participants referred to said continuous areas as opposed to single data points during their analysis for the task 1.

Task 11 indicates that cluster boundaries are better understood when they differ in color

Task 11 needed to be solved twice per participant, once for each approach. The task 10 before task 11 required participants to switch to a single-colored representation. They were asked to do so by choosing *country* as a sunburst hierarchy because there was only one country present in the dataset. It was anticipated that participants wish to switch back to the *IPC hierarchy* for the sunburst after completing the task. Notably, one of four participants still stayed on a single-colored *country* hierarchy kind. It appears that they did not understand where exactly the boundaries of a labeled area lie. They said "There are areas not occupied with big key terms" while they were describing areas on the largest cluster level. In fact, due to the nature of agglomerative clustering, every data point always belongs to some cluster. The area that was "not labeled" according to the participant's perception would have been the same color as the points around the nearest "labeled" area. Conceivably, it would have been easier for the participant to identify boundaries of the cluster correctly if they had perceived one color within the cluster but different colors outside the cluster.

One participant had selected *IPC classes* as the sunburst hierarchy for both approaches. While describing the areas, they circled with the mouse cursor around the boundaries of the clusters in a fairly precise way. Other two participants had *country* activated for one approach and *IPC classes* for another. One of them circled around cluster boundaries on a colorful representation with IPC classes, but not on a single-colored representation. This

might speak for the hypothesis. The other participant moved the cursor from cluster label to cluster label, which is why it is impossible to estimate their perception of the cluster limits.

We anticipated that the participants would make a connection between the content of an IPC class and cluster labels. For example, one could expect reasoning analogous to the following: "Areas with more blue points are overwhelmingly about chemistry, and the cluster about contact lens containers includes a group of blue points, so that means the containers are probably treated with or hold some chemical solution." Unfortunately, no participant voiced any indication that they matched the labels of clusters with the content of the IPC hierarchy. It is nevertheless possible that they would have made this connection if they had been having the chance to work with the prototype for a longer time. Participants themselves emphasized that they would normally spend significantly more time to familiarize themselves with the dataset, including the IPC classes that are used within the data.

General observations

During the study, there were quite a few instances when colors within the scatter plot changed depending on what node was currently selected in the sunburst. In every case, the mapping between colors and metadata values was directly obvious to the participants. We therefore conclude that the dynamic color assignment was understood.

Summary

Ultimately, it was confirmed that the changing of the colors of the points depending on the sunburst state is understood. As for whether it helps identify the clusters, this part of the hypothesis is more likely to be confirmed than refuted. Further experiments would be necessary to collect more evidence.

6.2.1.2. Hypothesis 2: Glyphs as indication of co-occurrence of multiple values per metadata attribute are understandable and clearly visible

There were two tasks for evaluating this hypothesis:

- Task 1. What IPC classes (on the section level, one letter) appear together often?
- Task 3. Choose one of three largest assignees. Does this institution collaborate a lot with others? If yes, are they other institutions or individuals?

Co-occurence of IPC classes

Intended solution

Our intended way of solving task 1 was to hover over each letter, one after another on the upper level of the IPC hierarchy. For each letter, the participants were supposed to roughly estimate how often the corresponding color appears on its own and with other

colors. One would also pay attention to the percentages in the breadcrumbs to support the estimation. The expected conclusion would be that all patents belong to the class G - "Physics", while classes A, B and C are assigned to about a third of the patents each. A majority of the patents belongs to two IPC classes and a smaller number to three classes. The most popular combination of two classes is G+A (red+green) occurring in 31% of patents.

Participants erroneously assume that choosing multiple sunburst nodes would show co-occurrences Two participants wrongly assumed that it was necessary to select multiple sunburst nodes simultaneously to display cooccurences. One participant glanced to the keyboard which was intentionally placed out of reach. They searched for a way of selecting multiple nodes, such as clicking on them while holding a control key. This erroneous attempt stemmed from the experience with desktop software where such mode of selection is popular, for example Windows Explorer or Microsoft Excel. One participant also used the right-click hoping to find a suitable menu option in the context menu.

In web applications, right-clicking and using key combinations are not conventionally used. From the above, we conclude that it was not immediately obvious to the participants that the prototype is a web application. This conclusion is supported by the fact that participants were slightly confused when asked to refresh the web page when technical problems required a reset of the current state.

Expert's perception of co-occurrence values was correct but occasionally incomplete The first expert's initial impression was that G (red) occurs most often and that G+A (red+green) is the prevalent combination. Then the participant examined color combinations by hovering the cursor over all one-letter IPC classes one after another. At the same time, they read percentage values in the breadcrumbs which confirmed their estimations. They realized that 100% the of data points belonged to IPC class G and guessed the reason for that correctly: "It was probably used in the query to produce the dataset". The impression that red+green appears most often was also confirmed as the participant saw the value of 31%. All in all, this solution matched our expectations completely.

In addition to the first expert, one more participant was able to discover that 100% of the data points belonged to IPC class G - "Physics". The remaining two experts nevertheless mentioned that the color of this class (red) occurs most often. The total number of participants who saw the percentage values for classes A, B and C is two. In the end, three experts were able to provide at least a partial solution for task 1.

One participant navigated into section A on the first level of the IPC hierarchy and explained co-occurrences between subclasses of A, such as A61 and A45. They however were unable to describe co-occurrences on the top level of the IPC hierarchy when prompted, even though the course of action would be exactly the same and just the IPC selection different.

Colors in glyphs mix when seen from far away In one case, the participant perceived purple color for glyphs composed of red and blue:

"On the first glance I thought it was purple and I could not find purple on the sunburst". We assume that the colors merged the same way blue, green and red diodes together can create a pixel of any color. The effect did not apply any longer after the participant had zoomed in and glyphs had become slightly bigger, which supports our assumption.

Co-occurence of assignees

Intended solution

The expected procedure for solving this task would be to navigate into the chosen assignee node by clicking on it. Then, the user would get an impression of how many points have just one single color corresponding to the selected assignee. If there were many multi-colored points that have multiple assignees, the participant would hover the mouse cursor over them and read the assignee entries in the detail view. This way it would be possible to say if collaborators were mostly institutions or individuals.

Experts' solutions

One participant understood *cooperation* from the wording of the task not as co-occurrences of assignees, but as organizations citing each other's work. As described in subsubsection 2.1.3.3, this understanding is incorrect. A citation purely indicates that the inventors are aware of existing inventions and still consider their patent novel.

All experts started the task by switching the sunburst hierarchy to *assignees*. Despite the large number of assignees in the hierarchy, it was obvious to all users that assignees were ordered in the descending order by the number of applications. One participant tried to detect co-occurrences by hovering over the tree top assignees in turn. They were looking for points that stay visible after switching to another assignee. In the end, all except one participant clicked the chosen assignee node, which happened to be Novartis AG in all cases, probably because it was the largest assignee. Then they hovered over multi-colored points and read the assignee information in the detail view for a number of patents.

One expert, however, did not realize that it was the intended way to solve the task: "I can hover here and I get the assignee information, but it surely wasn't meant like this." They wished for a list of all assignees the chosen assignee has ever cooperated with. According to them, only with such a list would they be able to see if the cooperations were mostly with organizations or individuals.

One more participant wished an alternative representation of assignee co-occurrences. They imagined a visualization in matrix form, possibly because they were familiar with such representation from the STN AnaVist software. They also imagined a sunburst hierarchy *assignee - assignee* in which the co-occurrences of assignees would be shown on the second level. The idea would not be applicable to IPC classes since it is not possible to display co-occurrences on different levels of IPC hierarchy clearly in a way that is consistent with current interaction techniques for switching between levels. Most importantly, in the current implementation of the sunburst child nodes indicate subgroups

and not co-occurrences, which means that one would be introducing contradictions into the mental model of the sunburst control.

Ultimately, three experts were able to complete the task and answer the question. Their conclusion was that Novartis AG does not cooperate much with other institutions. When they do, it is mostly with individuals. According to the experts, a characteristic feature of the US patent law is to enter the names of the inventors working for a company into the patent application as an assignee along with the the company itself. Two experts explained that those individual assignees are most probably inventors working for Novartis AG.

Summary

Over the course of the two tasks, it has been confirmed that glyphs allow users to easily distinguish data points with one value of a categorical attribute from the points with multiple values. When the goal is to provide a first impression about the distribution of values of metadata attributes, glyphs fit the purpose. For a more detailed quantitative analysis of co-occurrences other representation forms are more suitable, for example, a co-occurrence matrix or a bar chart.

Furthermore, not using glyphs would mean that each data point would only be assigned one color. Choosing to display only one value of a metadata attributes from a list of values would skew the perceived distribution of values because many values would not be displayed. Considering all of the above, we consider the hypothesis confirmed. Glyphs as they are used in our approach are an understandable indication of co-occurrence. However, they are not sufficient for a detailed co-occurrence analysis.

6.2.1.3. Hypothesis 3: The sunburst is suitable not only for hierarchical attributes, such as IPC classes, but also for arbitrary sets of categorical attributes

The task for evaluating this hypothesis was Task 5: “Compare in what IPC areas *Novartis AG* and *Johnson & Johnson Vision Care* are active”.

Intended solution

For this task, participants were expected to switch the sunburst hierarchy to *assignees - IPC codes*. Then the anticipated course of action would be to navigate into the assignee node by clicking on it, examine the distribution of the IPC classes, go back to the root of the hierarchy and do the same with the other assignee.

Switching the sunburst hierarchy kind

The previous task left the sunburst hierarchy set to *assignees* and all participants understood the need to switch hierarchy kind. This was the first time the participants needed to use two metadata attributes simultaneously in the sunburst. One participant first navigated to Novartis AG on the assignee level, then attempted to change the hierarchy to *IPC classes*, which brought invalid results due to a technical problem. Apparently, they expected that the currently selected hierarchy node persists after switching to another kind of hierarchy.

After they realized their misconception, they were able to choose the expected hierarchy kind successfully.

Navigation was problematic because of performance issues

Participants had a reliable grasp on navigation between the levels of the sunburst. All of them realized the need to navigate into a node if its child nodes are currently too small for convenient interaction.

The assignees mentioned in the task are ones in the top three according to the total number of patent applications. This is intentional because of the large number of assignees in the dataset (about 1100). Measures were taken to merge assignees with different spellings into one entry as described in subsection 5.1.3 "Parsing of metadata attributes". However, the resulting view still permitted working comfortably with approximately top 20 assignees only. The rest of the assignees occupied a sector of the circle too small to be selected reliably with the mouse.

As is, using assignees as the first level of the sunburst hierarchy resulted in a suboptimal experience for the participants because of performance problems. For some participants, it took time to learn to wait until the hover or click interactions were completely rendered by the prototype. When they did not immediately see the result of their actions, they moved the mouse cursor around and accidentally hovered over small assignees they did not recognize. After participants learned what response times to expect, all of them were able to successfully navigate to an assignee of choice and back to the root of the hierarchy. One expert said: "I thought you could type it in here somewhere" referring to the choice of an assignee.

Summary

All experts were able to describe the differences in the distribution of IPC classes for two chosen assignees. They compared the size of the same sunburst nodes not only on section level (e. g. G), but on the class level (e. g. G02) as well. One expert stated that the differences "can be seen well".

One conceivable solution to the challenge of showing many nodes on the same hierarchy level would be to enlarge the sunburst control to fit the whole screen. For hover interactions to be possible while still seeing the data points, one would move the enlarged sunburst control to a second monitor. An alternative solution would be to only distinguish between the top 10 to 20 assignees and aggregate the rest into the category "Others". The disadvantage of this approach is that it would be difficult to perceive the distribution of assignees by the number of applications, i. e. whether the technology domain is dominated by a few powerful companies or many smaller ones.

Until a solution is implemented, we would discourage from using the sunburst chart for metadata attributes with tens of different values. For the patent landscaping scenario, country and IPC class are attributes that result in a clear usable representation in a sunburst, while assignee is not suitable. The hypothesis is therefore considered partly confirmed.

6.2.1.4. Hypothesis 4: The interaction of all parts of the interface is understood

There were two tasks that could only be solved if the interaction between the sunburst view and the histogram view was clearly understood by the participant:

- Task 2. During what timespan does the IPC area G02C13 (Assembling; Repairing; Cleaning) actively develop?
- Task 4. Compare the timelines of applications for the assignees *Johnson & Johnson Vision Care* and *Bausch & Lomb*.

Furthermore, no view in the interface is so independent from the others that it could provide full benefits if taken alone. This means the remaining tasks also significantly contributed to the testing of this hypothesis.

Trends in development of an IPC class

Intended solution

This task was intended to be completed as following. First, the sunburst hierarchy has to be switched to "IPC classes". Second, one needs to interact with the node that represents G02C13 either by hovering over it or by clicking on it. In case of hovering, the number of applications for this subclass will be overlaid over the histogram view in the matching color. In case of clicking, only applications in this subclass will be shown in the histogram view. The participant would then be able to detect application peaks, time intervals with a significant number of applications and with little to no applications.

Interaction between sunburst and histogram is clear

All four participants preferred clicking on G02C13 instead of hovering over it. We suppose that clicking is a more widespread and intuitive interaction form than hovering and therefore participants were more comfortable with clicking. Importantly, every expert connected the change in the values of the histogram to their previous action of selecting the subclass. Accordingly, all four were able to describe the trend of G02C13's development. The years they named as a period of active development were consistent between participants.

Trends in development of an assignee

Participants' solutions

The solution of task 4 was very similar to the one of task 2, except the participants should change the sunburst hierarchy kind to *assignee*.

As described in the previous paragraph, during task 2 no participant has intentionally used the hover interaction on a sunburst node to see the number of applications per year for that node. Notably, by task 4 two of the participants became accustomed to the hover interaction and successfully used it to solve the task.

One participant navigated into the required assignee node, then back up and to the second assignee. While doing so, they expressed some frustration about not being able to see the timelines for both assignees at the same time: "One cannot remember all of this". They intended to perform a thorough year-per-year comparison, which is more effort than the task actually required.

In fact, this is where the advantage of the hover interaction becomes apparent as compared to clicking into the assignee node. Going one level back to switch to another assignee requires a change of context. A hover interaction, however, takes place at the root level of the sunburst hierarchy and does not require switching back and forth. It also takes less time so it is easier to keep the values of the time-series in short-term memory. Nevertheless, we agree that for detailed quantitative comparison it is necessary to display both time series simultaneously.

Interaction between sunburst and histogram is clear and adds value

In the end, all experts were able to complete the task. After they saw no new applications for Bausch & Lomb since 2009, one participant was able to put this fact in the context of their domain knowledge: "it stops then because the company does not exist anymore". One participant said that "it is often interesting to see this" referring to the evolution of the domain through the years. Another participant agreed that it was "by all means" useful to be able to examine the time dimension.

Tasks 2 and 4 allow us to confirm that the interdependence between the sunburst nodes and the time series values in histogram view was indeed clear to the participants. Our visualization approach is thus considered suitable for the *trend analysis* usage scenario.

Interactions between the sunburst itself and the breadcrumbs

It is impossible to say exactly how the breadcrumbs view was utilized without a camera or an eye-tracking system which we did not use. Nevertheless, reading can be recognized when the person whispers the words or says them aloud, or when the person moves the mouse to the text they are reading. It is especially interesting to know whether the experts took advantage of the full descriptions for the IPC classes and unshortened assignee names.

We can say confidently that the changes in the breadcrumbs view depending on the the actions in the sunburst were clear to all participants. For example, if they navigated into a sunburst node, they were reassured of the result of their action because the navigation path was shown with the breadcrumbs. Notably, for tasks involving IPC classes, the participants often read the descriptions for the upper level of the IPC hierarchy, but seldom for deeper levels. We attribute this to the fact that the experts were not familiar with the domain of contact lenses. They probably did not see the point in getting acquainted with the IPC classes from this domain, especially considering the short-term nature of the tasks in the study.

Providing the descriptions for IPC classes implements the principle of *recognition instead of recall* as described in subsubsection 5.2.3.2. The experts can easily browse through the nodes in the sunburst and recognize the thematic areas they search for. This feature was

also requested by the experts during the feedback meeting when the first iteration of the prototype was presented. Therefore, textual content along with the graphic breadcrumb nodes provides added value. For the above-mentioned reasons, we are convinced that the visual metaphor for the breadcrumbs view is chosen appropriately.

Interactions involving scatter plot and detail view

Task 3 (assignee co-occurrences, see subsubsection 6.2.1.2) and all tasks from part 2 of the evaluation relied on information only visible in the detail view. Except for cases of overlapping points (see subsubsection 6.2.1.5) or points obscured by cluster labels, the participants were able to retrieve detailed information for a chosen point confidently.

As with hover vs. click interaction on the sunburst (see the beginning of this subsection), some participants were initially more comfortable with clicking, even when they did not need the information in the detail view to persist (e. g. for comparison).

Summary

Summing it up, the participants grasped the interplay between sunburst and histogram, sunburst and scatter plot (see subsubsection 6.2.1.1), scatter plot and detail view (see above). Notably, the training in the beginning of the study only took 10 minutes. After the training the experts were able to reliably control the visualization. We therefore consider the hypothesis confirmed.

6.2.1.5. Hypothesis 5: Dynamic density of labels for points and clusters results in a readable and informative presentation

Intended for evaluating this hypothesis was task 6: “Navigate into the dataset with panning, zooming in and back out and assess the readability of labels in different levels of detail”. The task is self-explanatory.

One participant asked why a point had no label while they were on a middle detail level. As described in subsubsection 5.2.1.1 “Zooming”, adaptive density of labels was implemented for readability reasons. This fact, however, was not included into the introduction to the study for brevity’s sake.

Participants described the point labels as well-readable with just one exception. When the scatter plot was zoomed in to the maximal detail level (zoom factor 10), the top three relevant terms were shown per patent. In some cases the points were situated close enough for the labels to overlap and become unreadable. This happens when points are approximately at the same position horizontally and the labels are long. According to one expert, if two points are so close that they overlap, “it is even more important to know what the differences [between them] are”. The expert hoped to comprehend the difference based on the top three key terms.

The participants asked for the overlapping problem to be solved and suggested solutions themselves. For example, they suggested automatically moving overlapping points away from each other at the maximal zoom level. One participant asked if they could “play

around" with overlapping points, with the intention to pull them apart. Allowing users to manually move points would be a simple alternative to an automatic overlap removal algorithm. Another simple solution that takes almost no implementation effort would be just increasing the maximal allowed zoom level so that points stand further apart. One more alternative suggested by an expert is to bring a point into foreground when it is hovered over. For this idea to work as intended, a contrast background has to be added behind the active point to separate it from its neighbors and make the label readable.

Summary

No participants perceived that there were too many or too few labeled points, which is exactly the result we were aiming for with the heuristic for a proportion of shown labels.

Our observation did not reveal any readability problems except the mentioned overlapping issue. The hypothesis is considered confirmed.

6.2.1.6. Hypothesis 6: Cluster labels computed based on word2vec-based embeddings correspond to human understanding in a better way than those in TF-IDF-based embeddings

The task specifically aimed at evaluating this hypothesis was Task 11: "Briefly describe broad thematic areas in the dataset (big clusters). Evaluate the positions of the clusters relative to each other". This task is very reliant on the significance of cluster labels. Most importantly, all tasks from part two of the study implicitly assess this hypothesis.

Table 6.1.: Comparison of cluster key terms for both approaches. Contact lens dataset

№	Top 15 key terms per approach		Experts' description
	Semantic	Baseline	
1	**enzyme**, **cleaning**, **disinfecting**, **polyionic**, **polyionic material**, **salt**, **peroxide**, **surfactant**, **composition**, hydrogen peroxide, **polyoxyethylene**, **deposits**, **antimicrobial**, **polyanionic**, aqueous medium	**cleaning**, **polyionic**, **enzyme**, **surfactant**, **disinfecting**, **polyionic material**, medical device, **polyanionic**, medical, **composition**, **salt**, **antimicrobial**, **polyoxyethylene**, **peroxide**, **deposits**	-Cleaning -Enzymes, cleaning, hygiene, disinfection -Cleaning, disinfection -Enzymes for cleaning of lenses -Cleaning systems

Continued on next page

Table 6.1.: Comparison of cluster key terms for both approaches. Contact lens dataset

№	Top 15 key terms per approach		Experts' description
	Semantic	Baseline	
2	**prepolymer**, **divalent radical**, methylene pyrrolidone, macromonomer, **divalent**, **crosslinkable**, **vinylic**, **crosslinker**, **alk**, meth, parts weight, **hydrocarbon**, **denotes**, **compound**, polymerizable composition	**crosslinkable**, **prepolymer**, **compound**, **divalent radical**, **divalent**, **alk**, binder, radical carbon, **hydrocarbon**, **denotes**, **vinylic**, vinylic monomer, ink, binder polymer, **crosslinker**	-Materials the lens consists of, they are polymers -Materials -What the lens consists of -Substances, polymers the lens consists of -How the materials are produced
3	medical device, dye, medical, ink, article, **precursor**, **polymeric material**, reactive, **hybrid**, ophthalmic lens, cellulose, **hybrid contact**, **polymerizable composition**, anti, printing	**hybrid**, **polymerizable composition**, **hybrid contact**, **polymeric material**, **precursor**, siloxane monomer, collagen, core, aromatic, aromatic based, extracted, polymeric lens, macromer, rigid, composite	-Materials the lens consists of -Materials -Lens itself, its shape -Colorants -Products
4	**central zone**, **aberration**, **meridian**, **optical zone**, **vertical meridian**, **refractive power**, **central optical**, **model**, **inferior**, **spherical aberration**, **optic zone**, **segment**, **stabilization**, **lens design**, transition zone	**central zone**, **aberration**, **meridian**, **optical zone**, **refractive power**, **vertical meridian**, **central optical**, **model**, **optic zone**, **inferior**, multifocal contact, **spherical aberration**, **segment**, **lens design**, **stabilization**	-Optical power. When someone has astigmatism, how the lens should be bent, its form -Abberation is how you see -Vision corrections as known from an optician
5	**container**, **chamber**, **package**, **housing**, **cleaning**, **lens package**, **reservoir**, **holder**, **cap**, **lens holder**, **cup**, **lens cleaning**, shaft, insert, carrier	**container**, **chamber**, **package**, **housing**, **cleaning**, **lens package**, **holder**, **cap**, **reservoir**, mold part, **lens holder**, station, **lens cleaning**, **cup**, male	-Containers -Packaging -Where the lenses are stored -How packaging is, how lenses are stored, reservoirs etc.

Continued on next page

Table 6.1.: Comparison of cluster key terms for both approaches. Contact lens dataset

№	Top 15 key terms per approach		Experts' description
	Semantic	Baseline	
6	**iris section**, **iris**, **shade**, **color**, **limbal ring**, **starburst**, **limbal**, **colored**, **dots**, **cosmetic**, **colorant**, white, ophthalmic lens, beam, **colored contact**	**iris**, **iris section**, **shade**, **colorant**, pattern, **colored**, **cosmetic**, **limbal ring**, **color**, **starburst**, **limbal**, **colored contact**, **dots**, dot, indicator	-Colored lenses -How the iris is built -The eye, the iris as an organ
7	**substrate**, **signal**, **sensing**, **information**, **antenna**, circuit, sensors, circuitry, **data**, wireless, disposed substrate, **energy**, reader, electrodes, **ophthalmic device**	**ophthalmic device**, **substrate**, **energy**, **signal**, dynamic, beam, ophthalmic lens, **data**, **information**, **sensing**, insert, filter, display, **antenna**, reflected	-Not clear -Some physical methods -Antennas, circuits, sensors. Those are devices that measure properties of lenses -Sensors, measuring devices -The end product (device), but also electronical properties, energy

For both semantic and baseline (TF-IDF-based) approaches, the top level of hierarchical clustering consisted of 7 clusters. Table 6.1 shows cluster terms for both approaches and the way the experts described the content of the corresponding area (see subsection 5.1.7 for details on key term extraction). For 5 out of those 7 clusters (clusters №1, №2, №4, №5, №6), the key terms were very similar between the approaches. In fact, at least two of the top three terms were exactly the same for both approaches. In case of clusters №4, №5 and №6 even all three terms were the same.

Experts were often able to consistently identify same areas between different approaches

Predictably, after the participants identified and described large thematic areas for one approach, they were able to instantly recognize comparable areas for another approach. The experts' descriptions of those areas were consistent:

- Cluster №1 is about cleaning of contact lenses with enzymes.
- Cluster №2 contains materials contact lenses are made of.
- Cluster №4 is about optical properties of the lenses and how they assist vision correction.
- Cluster №5 describes the storage of lenses in suitable containers.

- Cluster №6 is about colored contact lenses, according to three experts. One expert was led astray by the term "iris" as they assumed it could only refer to the organ. Human body parts, however, are not patentable by themselves. In this case, the iris is discussed in the context of a colored contact lens. An artificial pattern on the lens covers the iris and sometimes tries to replicate the aesthetics of a real iris.

In contrast to the abovementioned five clusters, the two remaining ones (clusters №3, №7) lead to most incoherent descriptions. Cluster №3 includes fairly inhomogenious topics, which resulted in the top three terms being completely different between the approaches (medical device, dye, medical vs. hybrid, polymerizable composition, hybrid contact).

"Medical" is, in context of contact lenses, a very general term that did not contribute much to the understanding of the domain. The participants struggled with its interpretation during the evaluation of the semantic approach. By contrast, chemical terms from the baseline approach were specific enough for a better understanding. For both approaches, the cluster №2 labeled by additional chemical terminology (crosslinkable, prepolymer, compound) was very close to the cluster №3 and partly merged with it. Because of that, for the baseline approach some participants described both clusters №2 and №3 together as containing materials for producing contact lenses.

Experts appeared to have difficulties describing cluster №7 because they were not very familiar with the domain of electronics, in particular wearables. The term "substrate" in this case refers to the material the electronic circuits are placed upon . Most participants have a background in chemistry, where substrate refers to the chemical of interest that is being modified [70]. It is possible that the discrepancy in the understanding of this term led to some confusion. In the end, no expert described this area sufficiently well.

Relative importance of top key terms

The participants' ideas about the content of the clusters seemed to be influenced greatly by the top three terms per cluster that are always visible. After an initial assumption about the topic of the cluster was made, the remaining 12 of 15 top terms (shown on mouseover) seemed not to alter the first idea. This might be explained by the perceptual principles of the visual hierarchy.

Objectively speaking, the fourth top term does not differ much in relevance compared to the third top term. In fact, all 15 top terms enhance each other well and provide meaning for the user when considered together. However, the fourth and further key terms are only visible on demand, and are shown in a much smaller font size. Those are the trade-offs that we chose to make for the sake of readability. It is likely that because of that, the perceived relevance of the fourth and further top terms is much lower, and they are not taken into account by the experts as much.

If we were to increase the font size of the on-demand terms to match the top three, we would have to distribute them around the cluster to avoid overlaps and visual clutter. Depending on the length, some terms might even land outside the cluster boundaries. All of this would suggest to the users that the exact position of the terms inside a cluster matters, which is not the case here. A better solution remains to be found.

Experts' perceptions of key terms

Regarding the key terms themselves the experts had varied opinions. One participant struggled to see the added value in the extracted terms. According to them, there were many irrelevant terms among the ones extracted for clusters and single patents. They explained that, for example, "pharmaceutical composition" is a very common phrase in the domain of medicine, and "bottom portion" or "side walls" are used ubiquitously when describing devices of any kind.

We agree that such phrases have little explanatory power, but only in the case of a homogeneous dataset where every patent is about medicine or about devices, accordingly. In our case, however, the contact lens dataset contained a diverse range of topics. The example phrases mentioned by the expert really can make the difference between different thematic areas visible. It is quite possible that for a less diverse dataset, such terms would not be considered "relevant" by the TF-IDF algorithm. Since they appear in the most patents within the corpus, the IDF would lower their computed relevance. A more extensive stopword list can nevertheless be of great help for the quality of extracted terms.

Notably, we used TF-IDF as a baseline approach for the key term extraction for single patents (the approach for clusters is described in subsection 5.1.7). We experimented with the algorithm's parameters, but the potential for improvement is not exhausted yet. Moreover, a number of more advanced algorithms based on TF-IDF exist [66] [56]. So do key term extraction methods not based on TF-IDF like Rapid Automated Keyword Extraction (RAKE) [88] and TextRank [71]. It might be beneficial to test the above-mentioned algorithms on patent texts.

The expert mentioned above was also wondering about the differences between adjacent points with disparate key terms: "To really analyze this, I need to know why those two are so close, and why are the extracted terms then so different?" We strongly agree with the requirement for explainability. In this case, however, the interpretation is made difficult by the fact that we are observing a space after the dimensionality reduction. One conceivable reason for dissimilar terms for supposedly semantically close documents might be a lot of shared vocabulary, which nevertheless does not show up in extracted terms as it is too widespread within the document corpus. This leaves the most narrowly used terms to be extracted as relevant.

Overall, the experts' perception was that the key terms are more useful for bigger clusters, where they also describe more general concepts: "On the higher levels the terminology is more helpful, so that you can associate it with something". The key terms become less helpful as one focuses on the specifics of a small group of documents. Nevertheless, one expert found multiple levels of detail for hierarchical clustering a helpful feature: "it has the generated terminology on two [sic] different levels, this is good".

Suggestions for improvements

The above-mentioned participant was also happy to see the areas labeled at all because "in other software you have to write the titles yourself". To give users the ability to compose their own captions for arbitrary areas is a feature that was already mentioned by the experts during the mid-term presentation of the thesis. It would be immensely helpful

for a detailed analysis to be able to draw a boundary around an interesting area. This would follow the principle of *recognition over recall*. It is a user interface design pattern that requires significantly less cognitive effort from the user when implemented. In this case, it takes less mental effort for the user to process an area they labeled themselves versus when they are forced to recollect their description of the area every time. Should our prototype be developed further as a product, this is a functionality that might be worth implementing.

All participants wished for a search functionality based on document terms when confronted with tasks that required zooming into the landscape: "it is difficult. I have this information [from the task] and I can't search, but can only zoom, hover and read". They would expect to see highlighted areas where the queried term appears often: "it would show me where it [the requested term] is and jump there. Maybe one [document] will be here and one there. It would say, here on the upper-left side there is a hotspot".

One participant had an impression that the baseline approach resulted in better key terms for the clusters, but could not elaborate on their vague feeling.

Cluster terms overlap significantly between approaches

To examine whether the similarity of cluster terms between both approaches is pertinent only to the contact lens dataset, we compared the top 15 terms for other datasets used in this thesis: 3d printers, video codecs, hair dryers and diesel engines. In the overwhelming majority of the cases, we could determine a one-to-one correspondence between clusters for both approaches based on common themes suggested by overlapping key terms. The results can be seen in section A.5 with the identical terms emphasized in bold.

It can easily be seen that a greater part of cluster key terms are exactly the same. The hair dryer dataset is the one where the corresponding clusters were most difficult to define. We attribute this to a small sample size (ca. 350 documents), where influence of noise and of individual properties of the algorithms is unlikely to be compensated. In fact, the number of matching terms among the top 15 terms per cluster increases with the size of the dataset. Clearly, matching terms can only be a result of a term extraction on approximately the same subset of the data. This shows that both semantic and baseline approaches group the same patents in a similar way, at least at the most general level.

Considering all of the above, with regard to key terms none of the compared approaches provided a significant advantage over the other. The hypothesis is therefore refuted.

6.2.1.7. Hypothesis 7: Distances between points in word2vec-based embeddings correspond to human understanding in a better way than those in TF-IDF-based embeddings

In the same way as with subsubsection 6.2.1.6, all tasks from part 2 of the study are applicable to evaluate this hypothesis:

- Task 7. Find the area/areas with patents about colored contact lenses.
- Task 8. Find the area/areas with patents about contact lenses with electronic components ("smart" contact lenses)

- Task 9. Find the area/areas with patents about cleaning of contact lenses.
- Task 10. Find the area/areas with patents about ordering systems for contact lenses. They handle interaction with the client: diagnosis, ordering of lenses for example as a subscription, adjusting the prescription, etc.
- Task 11. Briefly describe broad thematic areas in the dataset (big clusters). Evaluate the positions of the clusters relative to each other.

Searching for large areas

Tasks 7 and 9 were designed to be solved in a similar way for comparability, with one task per approach. For those tasks, we wanted to evaluate the approaches independently of the patent's metadata, only based on the textual content. For that, the participants were asked to choose "Country" as a hierarchy level for the sunburst. The dataset consisted of only patents from the US patent office, so filtering and highlighting features were not accessible for that single value.

Our examination of relative cluster placement

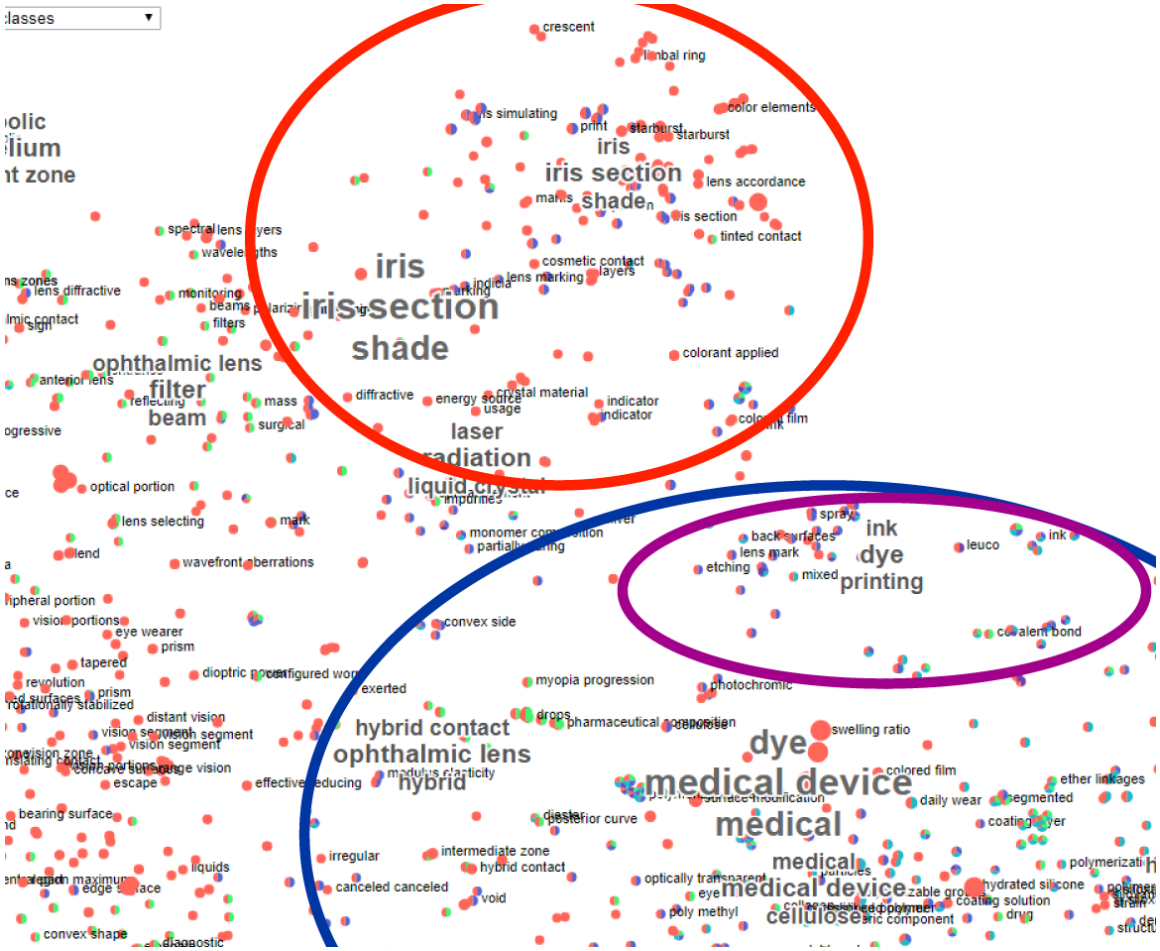
Both colored contact lenses and cleaning of contact lenses comprise their own large clusters (see clusters №6 and №1 in Table 6.1). Additionally, there are significant areas with related patents within other clusters. in Figure 6.1, we highlight the main large cluster of interest, the smaller relevant areas and the bigger cluster they are a part of.

For colored contact lenses, there exists an area focusing on ink for printing on colored contact lenses. For the semantic approach (see Figure 6.1(a), it lies inside cluster №3, which specializes on chemical substances for the production of lenses. The area about ink lies between the "chemical" cluster №3 and the cluster №6 with colored contact lenses.

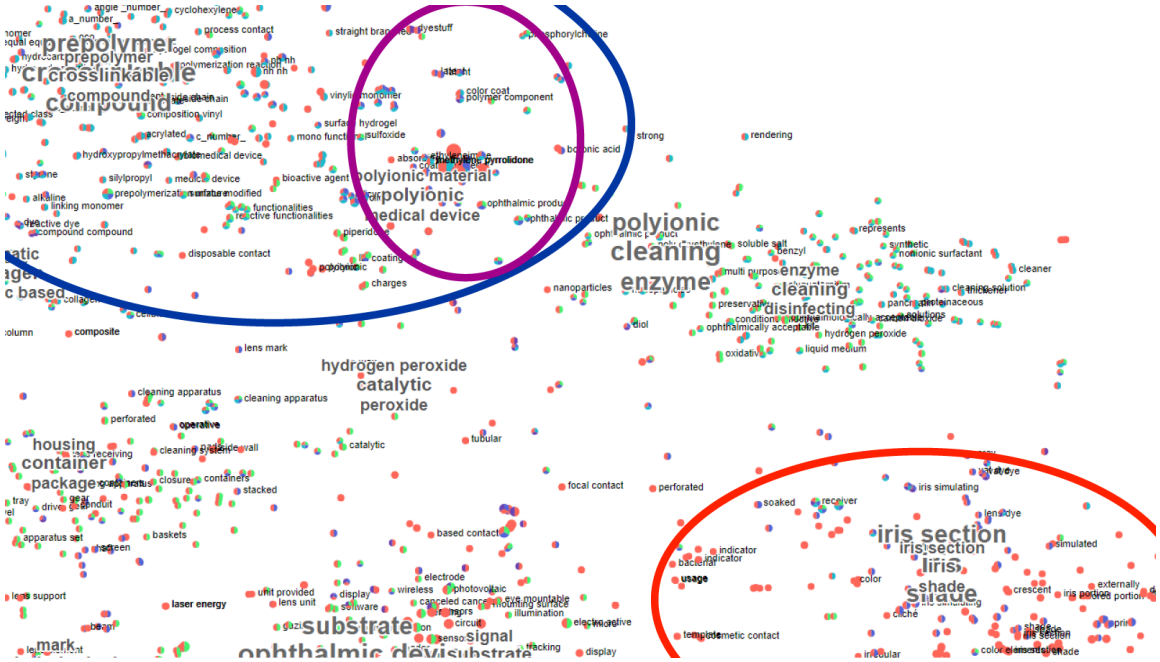
For the baseline approach, the cluster №3 (materials) and cluster №6 (colored contact lenses) are situated further away from each other and do not touch directly (see Figure 6.1(a). Nevertheless, the area about ink is at the edge of cluster №3 that is closest to cluster №6 .

For cleaning of contact lenses (cluster №1), a smaller related area is situated inside the cluster №5 about packaging and storage of lenses. It is predictable because containers for lenses often include a cleaning solution. Same as with task 7, the clusters produced by semantic approach are closer to each other (see Figure 6.2(a). For the baseline approach, the areas focused on cleaning and packaging are not adjacent (see Figure 6.2(b)). For both approaches, the area about containers with cleaning solutions is on the edge of cluster №5 that is closest to cluster №1.

In other words, between tasks 7 and 9 the semantic and baseline approaches behaved in a consistent way with regards to the relative placement of clusters. The semantic approach seems to place areas with strong thematic connections closer to each other. Notably, no other unrelated patents were situated between the two clusters of interest. The space between them was either empty or occupied with patents related to both clusters. This was not the case with the baseline approach, in which unrelated patents occupied the gap.



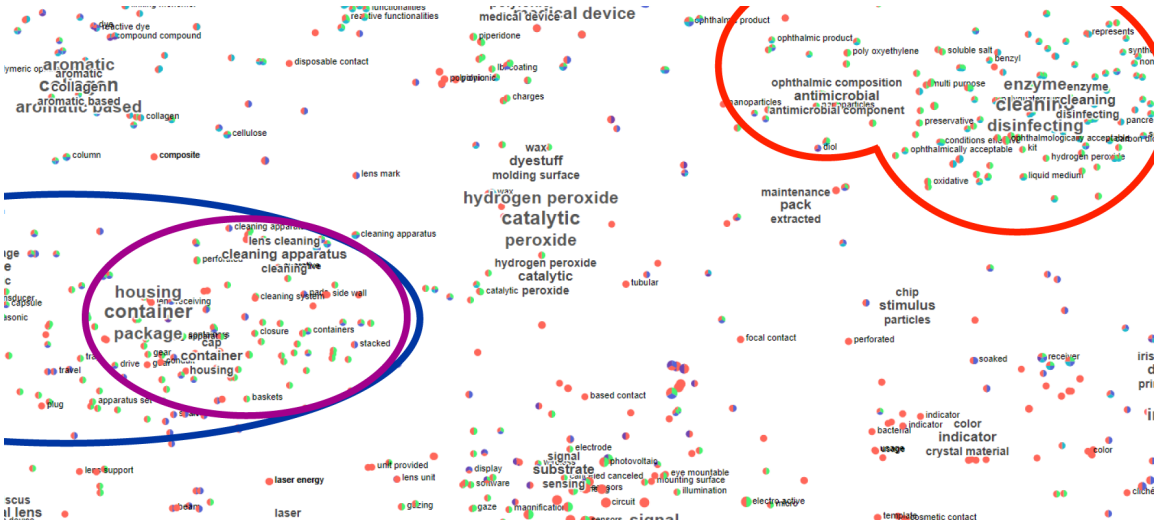
(a) Semantic approach



(b) Baseline approach

Figure 6.1.: Areas relevant for task 7. Colored contact lenses (red), materials for making lenses (blue), inks for printing on colored lenses (purple).

(a) Semantic approach



(b) Baseline approach

Figure 6.2.: Areas relevant for task 9. Cleaning of contact lenses (red), storage of lenses (blue), containers with cleaning solutions (purple).

One possible explanation for this phenomenon might be that the semantic representations result in more "continuous" areas where topics "flow" into one another. TF-IDF-based vectors, on the other hand, might produce more "interrupted" or "discrete" structures that are more likely to be rearranged during the dimension reduction process.

Intended solution

We assumed that the large clusters would be found fairly quickly based on their key terms. Then smaller related areas could be discovered by following the citation connections. The user is supposed to briefly hover over a number of patents within an area they consider definitely relevant. If most citation lines lead to one particular area outside of the current cluster, then it might be worthwhile to follow them and inspect the other area.

Summary

All participants found the requested large clusters for both tasks fairly quickly by reading the key terms for the large clusters. They then confirmed their assumptions by zooming into the selected cluster and reading terms, or in some cases, titles and abstracts of the patents situated there.

Two experts came up with the idea of using citations to find related areas. One of them even realized that many connections led to the same spot, but could not interpret this information further: "Even though with all those lines going to here and there I can see the other areas on the landscape, I still find it difficult to generate knowledge from that". This participant by chance examined only cited patents that were not directly thematically related to the region of interest. It is completely understandable that they could not see a direct connection. Ultimately, no expert could successfully identify the smaller related areas supplementary to the noticeable large clusters.

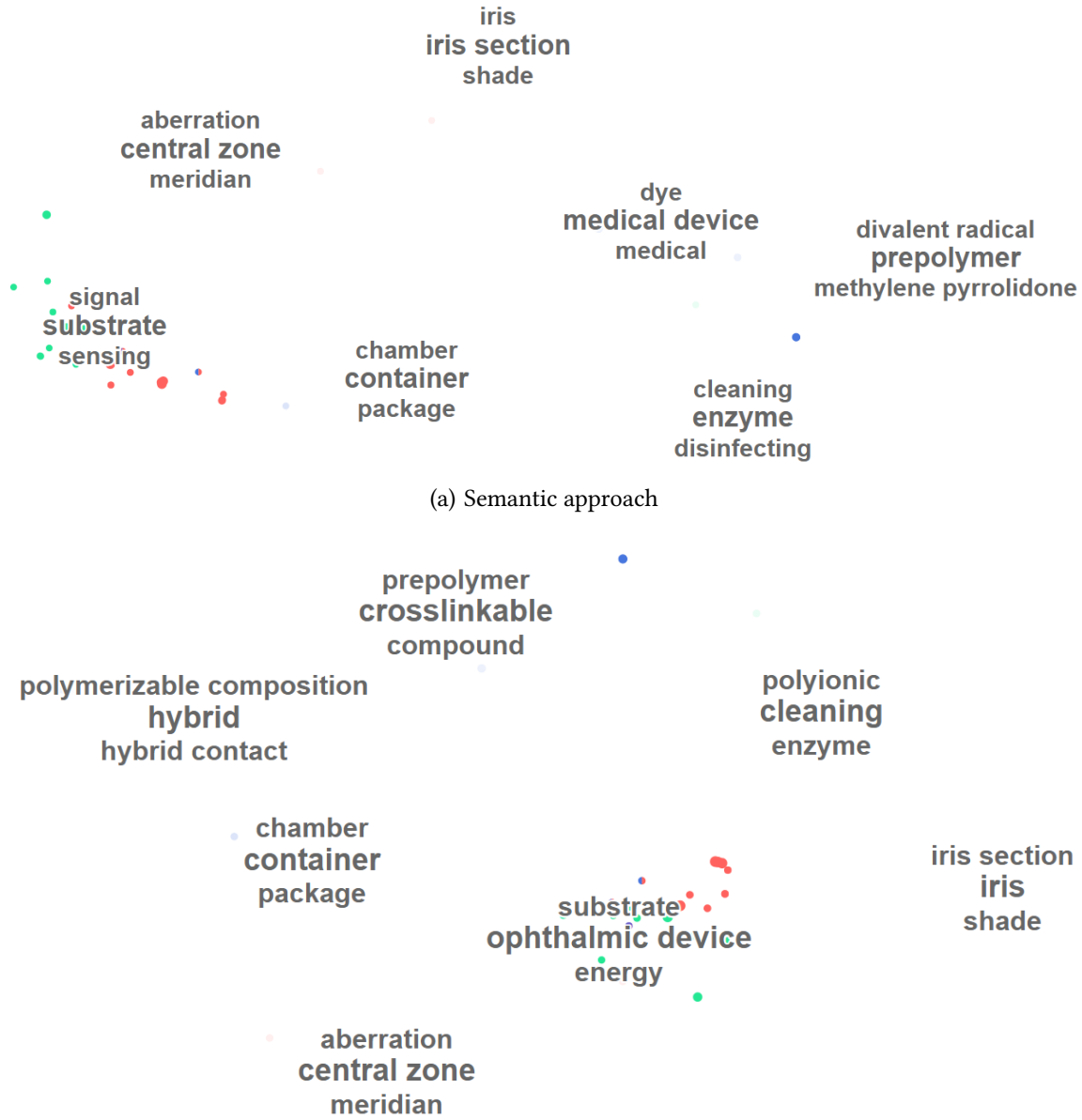
Searching for small areas

Same as with tasks 7 and 9, task 8 and task 10 were intended to be solved in a similar way for comparability. Here, using the filtering controls was allowed and encouraged.

Intended solution

For contact lenses with electronic components, one could safely assume that those patent applications are fairly recent. After restricting the timeline to the years after 2000, one can observe that the IPC class "H - electricity" grew in proportion from almost invisible 1.5% to 2.2%. Electricity is a plausible IPC section, so after selecting that "H" node, there are virtually no documents left except one densely populated spot (see Figures 6.3(a) and 6.3(b)). It approximately matches the cluster №7 with terms *substrate, signal, sensing* (for semantic approach) or *ophthalmic device, substrate, energy* (for baseline approach). The user could then try lifting the filters and checking whether the patents inside the area of interest, but from outside of IPC class "H" and those filed before 2000 are relevant, too.

Upon further inspection, with the semantic approach it should become clear that approximately one-fourth of cluster №7's area is mostly dedicated to various industrial automation



(a) Semantic approach

(b) Baseline approach

Figure 6.3.: The landscape as seen after the proposed sequence of actions for solving task 8. The landscape is restricted to years after 2000 and IPC class H - “electricity”.

systems. Those systems definitely include electronic components, while the lenses itself do not. The fact that they were grouped together with smart contact lenses was foreseeable and meaningful because of the shared vocabulary. The remaining three-fourths of the area comprise the intended answer to task 8 (see Figure 6.4(a)). For the baseline approach, the separation between patents pertaining to automation systems and patents about smart contact lenses is less clear (see Figure 6.4(b)). The patents irrelevant for this task intersperse cluster №7. This is evidence of the superiority of the semantic approach for this particular case.

Experts' solutions

The study showed that the intended sequence of steps for solving the task was too complicated, at least for our experimental setting. None of the participants chose to use any filtering options. Moreover, the sunburst hierarchy was switched to *country* after the previous task, which resulted in a monochrome view. No participant wished to switch back to *IPC class* for a colorful display. This might mean that for this task, the participants concentrated exclusively on reading cluster and document terms and did not see a need to involve other controls. Another explanation is also plausible: as the participants saw a single-colored representation, it did not occur to them to change it because they briefly forgot that it is possible. Maybe if they had a sunburst hierarchy unsuitable for this task already selected (i. e. *assignee*), they would have noticed the mismatch and would have tried to fix it.

The participants expressed the wish for a search feature because they would rather use it for such kind of task than browse around hoping to stumble upon the right area: "it's fairly difficult to find the right area if it [what you're looking for] is not labeled in the big regions".

Ultimately, three experts were able to identify cluster №7 as the relevant area. Unfortunately, they did not examine the area thoroughly enough to discover the irrelevant quarter of the documents. One expert started their search in the "chemical" area as they argued that polymers are the basis onto which the electronic components are placed: "the electronic components should be included in polymer materials somewhere". The participant browsed around until they stumbled upon the medium-sized cluster with the terms "signal, substrate, sensing", which they recognized as the area of interest.

Two other participants were steered to the correct answer by terms such as "signal, substrate, sensing" and "energy". They then confirmed their guess by reading some patent titles. The one expert who did find a wrong solution to the task was distracted by the plausible, but irrelevant term "medical device" in cluster №3. Our examination revealed that the term "device" in the above-mentioned area is not used to mean "appliance", but refers to the lens itself as a product for medical purposes.

Relative positions of clusters

For the semantic approach, one participant said "there are distinctly separated areas, it can be clearly seen". Another expert noted that "areas are separated more clearly" in the semantic approach, and that it looked "less noisy and more focused". Both of them

(a) Semantic approach

(b) Baseline approach

Figure 6.4.: Cluster №7 as seen in both approaches. The area relevant for task 8 is shown in red, the area about industrial automation systems is shown in blue, the cluster boundaries are shown in gray

also remarked that the clusters related to the chemical composition of contact lenses (see clusters №2 and №3 from Table 6.1) are supposed to be situated close to each other: “I think those two belong together”. One participant also noticed a difference in this area between the two approaches: “the polymer stuff was in two areas before and now it is in only one” (in this case, “before” refers to the baseline approach and “now” to the semantic approach).

Summary

Ultimately, the semantic approach seemed to reflect the structure of the domain in a better way. This, however, did not influence the experts’ strategy much, as the overall structure of the clusters was comparable. The proposed visualization approach itself mattered more than the method for placing the data points. When it comes to the specific tasks we offered, the hypothesis is likely to be confirmed.

Limitations of qualitative evaluation

It is important to note that the tasks created for the study cannot possibly evaluate the distances between the documents as a whole. When it comes to subjective perception, we can only witness the experts’ mental processes and their opinions. We can merely capture a small sample due to the limited number of tasks. A quantitative evaluation would be necessary to evaluate the placement of documents as a whole.

6.2.2. SUS

A German translation [85] of a SUS questionnaire was used after the first part of the study. The prototype received an average SUS score of 68.12 points, with a negligible difference of only one point between two tested approaches. It is important to remember that the SUS score is not a percentage, so it is necessary to consider the percentile when interpreting the value. According to [11], an application is considered “acceptable” at around 68-70 SUS points, which is also the average.

When participants explained their answers, they mostly named the technical imperfections of the prototype as reasons for lower scores. The performance on the contact lens dataset with about 2500 data points was not optimal and resulted in processing delays of up to two seconds, especially when hovering and clicking on the sunburst. It was important for us to provide an extensive dataset with sufficient thematic variation, so we willingly accepted the delays for that reason. The performance of the prototype was mentioned by the users: “I would use it, assuming it runs faster”, “The operation is not cumbersome, but difficult because of the delay. Because it is so simple, it should run smoothly”.

The experts confirmed that the prototype was easy to use: “It is easy because there are not infinitely many options to click on”, “I didn’t find it excessively difficult to use”, “It would not take long to train someone to use it”. They also missed features that one would expect from a finished piece of software: “What is implemented is coherent and conclusive, you could extend it nicely”, “It does not exploit all possibilities”.

While we are not proposing a commercially viable software product, we are conscious of the fact that it is difficult for users to evaluate prototypes with a limited number of features. Users are usually confronted with commercial software and therefore expect a comparable amount of development effort from every piece of software they encounter. Considering all of the above, we view the "average" SUS score as a success.

6.2.3. Questionnaire for comparing the approaches and the following mini-interview

If we consider differences in the answers between two approaches, they comprised 31% on average. 69% of the answers were identical for the same participant.

One participant felt that the baseline approach was "a bit better", but was not able to explain why. Other participants did not see a significant difference between the two approaches: "I am unable to say whether one or the other is better", "maybe I contradict myself [in the answers] compared to the other one, they are not that different".

All participants saw value in the proposed visualization approach as a whole, independently of the kind of document embeddings which is used. They emphasized that it is an easy way to get a general idea about the domain one is dealing with: "It is a fast possibility to be able to say what areas you are working with", "the topics are prepared how one would expect them to be". The expert compared the visual approach to patent landscaping with existing commercial solutions: "In other tools you can do an IPC analysis. This here is another approach with the same goal". They emphasized that with our proposed approach one must not heavily rely on knowledge of IPC classes.

One expert stated that "you can discover connections [with the prototype]". They expect that the prototype "can provide benefits in industry and research".

6.3. Discussion

6.3.1. The visualization approach and interaction metaphors

The prototype received an average SUS score which is considered acceptable, even though it lacked some features the participants expected or wished for. Their wishes were undoubtedly at least partly shaped by the participants' prior exposure to the patent landscaping tool STN Anavist, which is a commercial product and therefore cannot constitute a fair comparison to our proof-of-concept prototype. Overall, the participants' impression of the prototype was positive and they confirmed the need for such a tool.

Several relatively minor usability problems had been uncovered during the evaluation. First, panning and zooming start to noticeably lag when working with over 2000 patents, which resulted in performance problems. This was an inconvenience to the experts and affected their understanding of the current state of the sunburst. Another usability issue

that the participants remarked on were data points situated close to each other so that the labels overlap making the terms per patent unreadable.

Overall, the chosen visualization metaphors fit the task and were understood by the participants. The heuristic for the dynamic label density successfully provides a balance between text and whitespace. The dynamic mapping of colors depending on the state of the sunburst and glyphs as indicators of co-occurrence were quickly grasped by the participants. Just as well did the participants understand how the interconnected views affect each other's states. Considering the very short training the participants received, our visualization approach proved to be intuitive.

Thinking back to the research questions defined in the case study, we successfully found interactions techniques that are able to combine metadata of various types and semantic dimensions. The semantic space adds a dimension where one can find patterns via distributions of values of metadata attributes. For example, colors of the points likely contributed to the participants' perceptions of clusters.

The proposed visualization approach provides added value for various patent landscaping scenarios such as technology trend analysis. At the same time, there are no restrictions that would speak against the use of our approach for any kind of text documents characterized by metadata, for example, scientific publications.

The benefits of our proposed approach are especially evident when a brief overview of various thematic areas in the dataset is necessary. Detailed queries are better fulfilled using conventional tools such as textual search, bar charts and co-occurrence matrices.

6.3.2. Semantic embeddings versus TF-IDF embeddings

The think-aloud study resulted in inconclusive data with regards to the comparison of the two approaches. It has been shown that the approach itself only played a minor role in the evaluation process. The task-solving process was significantly more influenced by the interface of the visualization and its features.

According to some participants, the semantic embeddings resulted in a more intuitive relative placement of clusters. The clusters in the semantic approach were also separated in a better way according to our own examination and participants' perceptions. This might be attributed to the fact that semantic embeddings are dependent on the context of a word and therefore better capture similarities for synonyms, hypernyms, words used often in similar sentences, etc. However, the tasks we designed for the study could only possibly evaluate a subset of the documents with regard to their relative placement. Only a quantitative approach would be able to assess the positions of patents in visualization space as a whole. It is therefore impossible to draw a definitive conclusion at this point.

7. Conclusion

7.1. Summary

In this work, we investigated how to visually explore large document collections by employing semantics obtained from word embeddings of the document's textual content. We studied the problem for the task of patent landscaping as a case study. For that, with help of patent experts we studied the particularities of patent landscaping domain. We then proposed an interactive visualization approach that takes them into account.

We implemented a proof-of-concept interactive prototype. Similarities between documents are expressed through averaging weighted word embeddings of words in a document. The visualization makes the semantic space visible by reducing it to two dimensions with t-SNE. Additionally, multiple levels of detail are implemented via hierarchical clustering followed by a key term extraction. This helps make the local and global structures in the data visible, thereby supporting explainability of the semantic space.

Moreover, we incorporated metadata attributes of various types, for example, temporal, categorical and hierarchical, into the display through use of coordinated views. A zoomable scatter plot displays documents, while a sunburst and a histogram aggregate metadata values and serve to highlight and filter corresponding areas in the scatter plot. A detail view contributes to the exploration by providing maximal level of detail on demand. Taken as a whole, the user interface provided a way to discern patterns arising from the combination of semantically related clusters and the distributions of metadata values.

As a finishing part of the case study, we evaluated the prototype in a usability study with patent experts. We compared the word2vec-based document embeddings to TF-IDF vectors as sparse document representations.

7.2. Key results

The chosen interaction techniques proved to be consistent and intuitive. The study showed that the user interface of the prototype influenced the participants' perceptions significantly, while the way patents are situated and clustered did not play a major role. This is partly due to the fact that both approaches resulted in very similar extracted cluster key terms. The proportion of overlapping cluster key terms between both approaches increased with the size of a dataset. This can possibly be attributed to the greater influence of noise among local structures within the data for smaller datasets.

The semantic approach produced clusters that were better separated and placed more intuitively with regard to each other. The reason for this might be that semantic embeddings take the context of a given word, its synonyms, more specific or abstract words, etc. into account. This possibly results in a high-dimensional structure that is more cohesive and continuous as compared to the sparse TF-IDF-based representation.

The study results indicated that the combination of the semantic representations of documents' textual content and their metadata was understood by the participants and was likely helpful for finding clusters. Nevertheless, further research would be necessary to examine the mental processes involved in such exploration as it is a cognitively complex task.

The proposed visualization approach provides added value to the task of patent landscaping and can be applied to other document exploration tasks.

7.3. Future work

In this section we provide an outlook on the possible improvements of our approach, both general and restricted to the domain of patent landscaping.

7.3.1. Improvements independent of the patent domain

Patent landscaping depends heavily on the input dataset. [3] proposes a neural-network-based approach that expands the given seed dataset by following its citations *outwards*. The model they developed then prunes the patents that are not directly relevant to the seed's topic. This results in a more complete dataset because it now contains related and relevant documents that would have been omitted otherwise. Their approach would be extremely useful as a part of the data acquisition phase before our data processing pipeline.

The data processing for our visualization at the moment involves one manual step that is very influential for the result. It is the selection of suitable cut-off values for the three detail levels of hierarchical clustering. A single optimal clustering does not exist due to the subjective human perception. This means a heuristic must be introduced that would help find advantageous number and size of the clusters and possibly even the number of levels of detail depending on the size of the dataset.

Our approach addresses the challenge of visual scalability as it allows the users to explore hundreds to thousands of documents simultaneously. However, the computing power available to us was not sufficient for a smooth operation when showing ca. 2600 documents simultaneously. For large document collections, it would be immensely advantageous to allow a graceful degradation of the functionality for a fluid performance. A balance must be found between preserving the functionality and preserving an adequate response time.

Composing a document embedding out of word2vec embeddings is no longer state-of-the art. There has been a number of promising approaches for context-sensitive word embeddings or document embeddings such as paragraph2vec [61], ELMo [81], BERT [32]

etc. We chose word2vec as a simple standard approach which proved to be successful. Moreover, developing a new embedding method or training a model specifically for the patent domain based on an existing approach was not in the scope of this work. However, it might be worthwhile to compare different word and document embedding methods. Moreover, embedding different parts of the text document separately (in our case, patent's claims and sections of description) might provide interesting insights.

Considering the key term extraction for single documents, a number of advanced algorithms based on TF-IDF exist [66] [56]. Moreover, there are key term extraction methods not based on TF-IDF, such as RAKE [88] and TextRank [71]. It might be worthwhile to compare those algorithms to each other and to our straightforward implementation of TF-IDF.

Our experiments showed that t-SNE is the most suitable dimension reduction method at the moment. Its result, however, is dependent on the parameters of the algorithm, especially the perplexity. With a change in perplexity local structures within the data change their shape. It might prove valuable to let the user dynamically vary the perplexity to see how the cluster shapes change. It would, however, require a waiting process because the clusters and their key terms have to be generated anew.

We briefly attempted to derive general key terms from specific ones with unsatisfactory results. If one were to determine the exact meaning of a key term out of a multitude of its contextual meanings (synset detection), it would open a possibility to reliably augment cluster terms with hypernyms.

7.3.2. Patent-specific improvements

It came to our attention during the interviews with patent experts that added value patent databases exist. Patent texts in them are rewritten in a concise way by trained professionals. We assume that a semantic approach such as ours would perform significantly better on the data derived from such added value databases.

Inventors usually play a smaller role in the patent landscape than institutions. It could be useful to automatically detect private persons and companies and possibly hide single inventors for a less cluttered overview. A special case would be when a patent belongs only to physical persons without association with any institution. One should not hide all inventors to avoid showing patents without any single assignee. Alternatively, one could aggregate such inventors into one group called "Miscellaneous" or "Others".

There are certain patent properties that we did not take into account. One example is the kind code which distinguishes between application, grant, search report, correction, etc. and is build differently for each country. Citations can also be of different types. It might be profitable for patent experts to be able to explore this information in addition to the information we have already present.

Bibliography

[1] *3D Printing: Technology Insight Report*. Tech. rep. Gridlogics Technologies Pvt Ltd, 2014, p. 4.

[2] J. Abello et al. "A Modular Degree-of-Interest Specification for the Visual Analysis of Large Dynamic Networks". In: *IEEE Transactions on Visualization and Computer Graphics* 20.3 (Mar. 2014), pp. 337–350. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.109.

[3] Aaron Abood and Dave Feltenberger. "Automated patent landscaping". In: *Artificial Intelligence and Law* 26.2 (2018), pp. 103–125. ISSN: 15728382. DOI: 10.1007/s10506-018-9222-4. URL: https://doi.org/10.1007/s10506-018-9222-4.

[4] *About Wordnet*. 2010. URL: https://wordnet.princeton.edu/ (visited on 05/04/2019).

[5] Jay Alammar. *The Illustrated Word2vec*. Mar. 2019. URL: https://jalammar.github.io/illustrated-word2vec/ (visited on 04/03/2019).

[6] *Altair: Declarative Visualization in Python*. 2018. URL: https://altair-viz.github.io/.

[7] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings". In: *ICLR* (2017), pp. 1–16. ISSN: 0022-1899. DOI: 10.1016/B978-0-12-401688-0.00001-X.

[8] Vasco Asturiano. *Zoomable Sunburst with Labels - bl.ocks.org*. URL: https://bl.ocks.org/vasturiano/12da9071095fbd4df434e60d52d2d58d (visited on 05/12/2019).

[9] *Axiis : Data Visualization Framework*. 2009. URL: http://www.axiis.org/.

[10] Rafael E. Banchs. "A Comparative Evaluation of 2D And 3D Visual Exploration of Document Search Results". In: *Information Retrieval Technology*. Ed. by Azizah Jaafar et al. Cham: Springer International Publishing, 2014, pp. 100–111. ISBN: 978-3-319-12844-3.

[11] Aaron Bangor, Philip Kortum, and James Miller. "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale". In: *Journal of Usability Studies* 4.3 (2009), pp. 114–123. URL: http://uxpajournal.org/wp-content/uploads/sites/8/pdf/JUS%7B%5C_%7DBangor%7B%5C_%7DMay2009.pdf.

[12] Patrick Baudisch et al. "Keeping things in context: a comparative evaluation of focus plus context screens, overviews, and zooming". In: *CHI '02 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 4. 2002, pp. 259–266. ISBN: 1581134533. DOI: 10.1145/503376.503423. URL: http://dl.acm.org/citation.cfm?id=503423.

[13] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[14] Tim Bock. *What is a Dendrogram? How to use Dendrograms | Displayr*. URL: https://www.displayr.com/what-is-dendrogram/ (visited on 05/21/2019).

[15] *Bokeh*. 2018. URL: https://bokeh.pydata.org/en/latest/.

[16] M. Ted Boren and Judith Ramey. "Thinking aloud: Reconciling theory and practice". In: *IEEE Transactions on Professional Communication* 43.3 (2000), pp. 261–278. ISSN: 03611434. DOI: 10.1109/47.867942.

[17] Mike Bostock. *D3.js - Data-Driven Documents*. 2019. URL: https://d3js.org/.

[18] Mike Bostock. *Scatterplot Matrix Brushing - bl.ocks.org*. 2019. URL: https://bl.ocks.org/mbostock/4063663 (visited on 05/24/2019).

[19] Kevin W. Boyack, Brian N. Wylie, and George S. Davidson. "Domain visualization using VxInsight® for science and technology management". In: *Journal of the American Society for Information Science and Technology* (2002). ISSN: 15322882. DOI: 10.1002/asi.10066.

[20] *BreadCrumb Control for IOS 9 (Swift) for iOS - Cocoa Controls*. URL: https://www.cocoacontrols.com/controls/breadcrumb-control-for-ios-9-swift (visited on 05/12/2019).

[21] John Brooke. "SUS - A quick and dirty usability scale". In: *Usability Evaluation in Industry* (1996), pp. 4–7. ISSN: 1097-0193. DOI: 10.1002/hbm.20701.

[22] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. "Readings in information visualization: using vision to think". In: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[23] Sheelagh Carpendale. "Evaluating information visualizations". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4950 LNCS (2008), pp. 19–45. ISSN: 03029743. DOI: 10.1007/978-3-540-70956-5_2.

[24] C. Chen. "Searching for intellectual turning points: Progressive knowledge domain visualization". In: *Proceedings of the National Academy of Sciences* 101.Supplement 1 (2004), pp. 5303–5310. ISSN: 0027-8424. DOI: 10.1073/pnas.0307513100. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0307513100.

[25] Chaomei Chen. "Top 10 unsolved information visualization problems". In: *IEEE Computer Graphics and Applications* 25.4 (2005), pp. 12–16. ISSN: 02721716. DOI: 10.1109/MCG.2005.91. arXiv: arXiv:1011.1669v3.

[26] J. P. Chin, V. A. Diehl, and L. K. Norman. "Development of an instrument measuring user satisfaction of the human-computer interface". In: January (2003), pp. 213–218. ISSN: 1098-6596. DOI: 10.1145/57167.57203. arXiv: arXiv:1011.1669v3.

[27] J. Choo et al. “UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 1992–2001. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.212.

[28] Adam Cohen. *FuzzyWuzzy: Fuzzy String Matching in Python - ChairNerd*. 2011. URL: https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/ (visited on 05/20/2019).

[29] *Contact Lenses: Technology Insight Report*. Tech. rep. Gridlogics Technologies Pvt Ltd, 2014, p. 5.

[30] Edilson A. Corrêa, Vanessa Queiroz Marinho, and Leandro Borges dos Santos. “NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis”. In: (2017), pp. 611–615. arXiv: 1704.02263. URL: http://arxiv.org/abs/1704.02263.

[31] Christopher A. Cotropia, Mark A. Lemley, and Bhaven Sampat. “Do applicant patent citations matter?” In: *Research Policy* 42.4 (2013), pp. 844–854. ISSN: 00487333. DOI: 10.1016/j.respol.2013.01.003. URL: http://dx.doi.org/10.1016/j.respol.2013.01.003.

[32] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[33] S George Djorgovski et al. “Immersive and Collaborative Data Visualization and Analytics Using Virtual Reality”. In: *AGU Fall Meeting Abstracts*. 2018.

[34] Wenwen Dou et al. “ParallelTopics: A probabilistic approach to exploring document collections”. In: *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings* (2011), pp. 231–240. DOI: 10.1109/VAST.2011.6102461.

[35] K Anders Ericsson and Herbert A Simon. *Protocol analysis: Verbal reports as data*. the MIT Press, 1984.

[36] K. Anders Ericsson and Herbert A. Simon. “Verbal reports as data”. In: *Psychological Review* 87.3 (1980), pp. 215–251. ISSN: 0033295X. DOI: 10.1037/0033-295X.87.3.215. URL: http://s3.amazonaws.com/academia.edu.documents/37092743/Ericsson-Simon80.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA%7B%5C&%7DExpires=1475089614%7B%5C&%7DSignature=2UVSlQ8f3qRxRJXYi0mzspm3C4s%7B%5C%%7D3D%7B%5C&%7Dresponse-content-disposition=inline%7B%5C%%7D3B%20filename%7B%5C%%7D3DEricsson-Simon80.pdf.

[37] Sara Irina Fabrikant. “Evaluating the Usability of the Scale Metaphor for Querying Semantic Spaces”. In: (2007), pp. 156–172. DOI: 10.1007/3-540-45424-1_11.

[38] Paolo Federico et al. “A Survey on Visual Approaches for Analyzing Scientific Literature and Patents”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (2017), pp. 2179–2198. ISSN: 10772626. DOI: 10.1109/TVCG.2016.2610422.

[39] Eugene Garfield. "Historiographic mapping of knowledge domains literature". In: *Journal of Information Science* 30.2 (2004), pp. 119–145. ISSN: 01655515. DOI: 10.1177/0165551504042802.

[40] Mark Giereth et al. "Web based visual exploration of patent information". In: *Proceedings of the International Conference on Information Visualisation*. 2007. ISBN: 0-7695-2900-3. DOI: 10.1109/IV.2007.141.

[41] *Graphene: The worldwide patent landscape in 2015*. Tech. rep. UK Intellectual Property Office Infromatics Team, 2015. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment%7B%5C_%7Ddata/file/470918/Graphene%7B%5C_%7D-%7B%5C_%7Dthe%7B%5C_%7Dworldwide%7B%5C_%7Dpatent%7B%5C_%7Dlandscape%7B%5C_%7Din%7B%5C_%7D2015.pdf.

[42] Brynjar Gretarsson et al. "TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3 (Feb. 2012). DOI: 10.1145/2089094.2089099.

[43] Jacob Gube. *Breadcrumbs In Web Design: Examples And Best Practices — Smashing Magazine*. 2019. URL: https://www.smashingmagazine.com/2009/03/breadcrumbs-in-web-design-examples-and-best-practices/ (visited on 05/12/2019).

[44] Hayet Hadjar et al. "WebVR based Interactive Visualization of Open Health Data". In: *Proceedings of the 2nd International Conference on Web Studies - WS.2 2018*. New York, New York, USA: ACM Press, 2018, pp. 56–63. ISBN: 9781450364386. DOI: 10.1145/3240431.3240442. URL: http://dl.acm.org/citation.cfm?doid=3240431.3240442.

[45] Marti A Hearst. "Modern Information Retrieval, chapter 10. User Interfaces and Visualization". In: *Addison Wesley Longman* (1999). URL: http://people.ischool.berkeley.edu/~hearst/irbook/10/node3.html#SECTION00122000000000000000.

[46] Dominik Herr et al. "Visual exploration of patent collections with IPC clouds". In: *CEUR Workshop Proceedings* 1292.IPaMin (2014). ISSN: 16130073. DOI: 10.5121/ijcis.2012.2406.

[47] E. Hetzler and A. Turner. "Analysis experiences using information visualization". In: *IEEE Computer Graphics and Applications* 24.5 (Sept. 2004), pp. 22–26. ISSN: 0272-1716. DOI: 10.1109/MCG.2004.22.

[48] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417.

[49] *IPC Publication*. 2018. URL: https://www.wipo.int/classifications/ipc/ipcpub/ (visited on 05/12/2019).

[50] *IPython*. 2019. URL: https://ipython.org/.

[51] Xinyi Jiang and Jiawan Zhang. "A text visualization method for cross-domain research topic mining". In: *Journal of Visualization* 19.3 (2016), pp. 561–576. ISSN: 18758975. DOI: 10.1007/s12650-015-0323-9.

[52] Daniel K N Johnson and Matthew Whitehead. "A Tool for Visualizing and Exploring Relationships among Cancer-Related Patents". In: *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference* (2017), pp. 235–238. URL: http://cs.coloradocollege.edu/%7B~%7Dmwhitehead/CancerMoonshot/documents/iaai.pdf.

[53] Ian Johnson. *d3 workshop - bl.ocks.org*. 2018. URL: http://bl.ocks.org/enjalot/6641917 (visited on 05/12/2019).

[54] *JupyterLab*. 2018. URL: https://jupyterlab.readthedocs.io/en/stable/.

[55] *Keras*. URL: https://keras.io/.

[56] Su Nam Kim, Timothy Baldwin, and Min-yen Kan. "An Unsupervised Approach to Domain-Specific Term Extraction". In: *Proceedings of the Australasian Language Technology Association Workshop* 2 (2009), pp. 94–98.

[57] Emiel Krahmer and Nicole Ummelen. "Thinking about thinking aloud: A comparison of two verbal protocols for usability testing". In: *IEEE Transactions on Professional Communication* 47.2 (2004), pp. 105–117. ISSN: 03611434. DOI: 10.1109/TPC.2004.828205.

[58] Bruno Latour et al. "'The whole is always smaller than its parts' - a digital test of Gabriel Tardes' monads". In: *British Journal of Sociology* 63.4 (2012), pp. 590–615. ISSN: 00071315. DOI: 10.1111/j.1468-4446.2012.01428.x.

[59] Page Laubheimer. *Beyond the NPS: Measuring Perceived Usability with the SUS, NASA-TLX, and the Single Ease Question After Tasks and Usability Tests*. 2018. URL: https://www.nngroup.com/articles/measuring-perceived-usability/ (visited on 03/26/2019).

[60] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE Laurens". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. ISSN: 02624079. DOI: 10.1007/s10479-011-0841-3. arXiv: 1307.1662.

[61] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: (May 2014). arXiv: 1405.4053. URL: http://arxiv.org/abs/1405.4053.

[62] Timothy C. Lethbridge, Susan Elliott Sim, and Janice Singer. *Studying software engineers: Data collection techniques for software field studies*. Vol. 10. 3. 2005, pp. 311–341. ISBN: 1066400512. DOI: 10.1007/s10664-005-1290-x.

[63] Mark Levene. *An introduction to search engines and web navigation*. John Wiley & Sons, 2011.

[64] V. I. Levenshtein. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10 (Feb. 1966), p. 707.

[65] James R. Lewis. "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. Boca Raton, FL: Human Factors Group". In: *IBM Technical Report* 54.1 (1993), p. 786.

[66] Feifan Liu et al. "Unsupervised approaches for automatic keyword extraction using meeting transcripts". In: June (2010), p. 620. DOI: 10.3115/1620754.1620845.

[67] Marco Lui and Timothy Baldwin. "Cross-domain feature selection for language identification". In: *Proceedings ofthe 5th International Joint Conference on Natural Language Processing* 1967 (2011), pp. 553–561. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.308.4653.

[68] Marco Lui and Timothy Baldwin. "langid.py: An off-the-shelf language identification tool". In: *Proceedings ofthe 50th Annual Meeting ofthe Association for Computational Linguistics* July (2012), pp. 25–30. URL: https://github.com/saffsd/langid.py.

[69] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (Feb. 2018). arXiv: 1802.03426. URL: http://arxiv.org/abs/1802.03426.

[70] Alan D McNaught and A Wilkinson. *Compendium of chemical terminology*. Vol. 1669. Blackwell Science Oxford, 1997. DOI: 10.1351/goldbook.S06082.

[71] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Texts". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, Barcelona, Spain*. Vol. 45. 4. 2004. DOI: 10.1016/0305-0491(73)90144-2.

[72] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: (2013), pp. 1–9. ISSN: 10495258. DOI: 10.1162/jmlr.2003.3.4-5.951. arXiv: 1310.4546. URL: http://arxiv.org/abs/1310.4546.

[73] David Modjeska. *Navigation in Electronic Worlds: Research Review for Depth Oral Exam*. Tech. rep. Toronto, Computer Systems Research Group, University of Toronto, 1997.

[74] El Moatez Billah NAGOUDI, Jérémy Ferrero, and Didier Schwab. "LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting". In: August (2018), pp. 134–138. DOI: 10.18653/v1/s17-2017.

[75] R. Nakazawa, T. Itoh, and T. Saito. "A Visualization of Research Papers Based on the Topics and Citation Network". In: *2015 19th International Conference on Information Visualisation*. July 2015, pp. 283–289. DOI: 10.1109/iV.2015.58.

[76] Jakob Nielsen. *2D is Better Than 3D*. 1998. URL: https://www.nngroup.com/articles/2d-is-better-than-3d/ (visited on 04/04/2019).

[77] Jakob Nielsen. "Enhancing the Explanatory Power of Usability Heuristics". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '94. Boston, Massachusetts, USA: ACM, 1994, pp. 152–158. ISBN: 0-89791-650-6. DOI: 10.1145/191666.191729. URL: http://doi.acm.org/10.1145/191666.191729.

[78] Jakob Nielsen. *Why You Only Need to Test with 5 Users*. 2000. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (visited on 05/12/2019).

[79] *NumPy*. 2019. URL: https://www.numpy.org/index.html.

[80] Ann A. O'Connell, Ingwer. Borg, and Patrick Groenen. "Modern Multidimensional Scaling: Theory and Applications". In: *Journal of the American Statistical Association* 94.445 (2006), p. 338. ISSN: 01621459. DOI: 10.2307/2669710.

[81] Matthew E. Peters et al. "Deep contextualized word representations". In: (2018). arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365.

[82] Steve Portigal. "Interviewing Users: How to Uncover Compelling Insights". In: *Rosenfeld Media* (2013).

[83] *Python*. 2019. URL: https://www.python.org/.

[84] Robert R. Sokal and F Rohlf. "Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. Taxon 11: 33-40". In: *Taxon* 11 (Feb. 1962), pp. 33–40. DOI: 10.2307/1217208.

[85] Matthias Rauer. *Quantitative Usablility-Analysen mit der System Usability Scale (SUS) - Nachrichten, Tipps & Anleitungen für Agile, Entwicklung, Atlassian-Software (JIRA, Confluence, Bitbucket, ...) und Google Cloud*. 2011. URL: https://blog.seibert-media.net/blog/2011/04/11/usablility-analysen-system-usability-scale-sus/ (visited on 05/21/2019).

[86] Severino Ribecca. *Sunburst Diagram - Learn about this chart and tools to create it*. 2019. URL: https://datavizcatalogue.com/methods/sunburst%7B%5C_%7Ddiagram.html (visited on 03/28/2019).

[87] Colin Robson. "Real world research 2nd edition: A resource for social scientists and practitioner-researchers". In: *Malden: BLACKWELL Publishing* (2002).

[88] Stuart Rose et al. "Automatic Keyword Extraction from Individual Documents". In: *Text Mining*. John Wiley & Sons, Ltd, 2010. Chap. 1, pp. 1–20. ISBN: 9780470689646. DOI: 10.1002/9780470689646.ch1. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470689646.ch1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470689646.ch1.

[89] *Rug plots in the margins — geom_rug • ggplot2*. URL: https://ggplot2.tidyverse.org/reference/geom%7B%5C_%7Drug.html (visited on 05/12/2019).

[90] Per Runeson and Martin Höst. "Guidelines for conducting and reporting case study research in software engineering". In: *Empirical Software Engineering* 14.2 (Apr. 2009), pp. 131–164. ISSN: 13823256. DOI: 10.1007/s10664-008-9102-8. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://link.springer.com/10.1007/s10664-008-9102-8.

[91] Laura Ruotsalainen. *Data mining tools for technology and competitive intelligence*. 2451. 2008, pp. 1–64. ISBN: 9789513872410.

[92] *SciPy*. 2019. URL: https://www.scipy.org/.

[93] B Shneiderman. "The eyes have it: a task by data type taxonomy for information visualizations. Proceedings of IEEE Symposium on Visual Languages". In: (1996), pp. 336–343. ISSN: 1049-2615. DOI: 10.1109/VL.1996.545307. arXiv: arXiv:1011.1669v3. URL: http://portal.acm.org/citation.cfm?id=832277.834354.

[94] Ben Shneiderman. "Tree Visualization with Tree-maps: 2-d Space-filling Approach". In: *ACM Trans. Graph.* 11.1 (Jan. 1992), pp. 92–99. ISSN: 0730-0301. DOI: 10.1145/102377.115768. URL: http://doi.acm.org/10.1145/102377.115768.

[95] Herbert A. Simon and Allen Newell. "Human problem solving: The state of the theory in 1970." In: *American Psychologist* 26.2 (2006), pp. 145–159. ISSN: 0003-066X. DOI: 10.1037/h0030806. arXiv: arXiv:0910.0734v1.

[96] A. Skupin. "The world of geography: Visualizing a knowledge domain with cartographic means". In: *Proceedings of the National Academy of Sciences* 101.Supplement 1 (2004), pp. 5274–5278. ISSN: 0027-8424. DOI: 10.1073/pnas.0307654100. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0307654100.

[97] André Skupin. "A Cartographic Approach to Visualizing Conference Abstracts". In: *IEEE Computer Graphics and Applications* 22.1 (2002), pp. 50–58. ISSN: 02721716. DOI: 10.1109/38.974518.

[98] André Skupin, Joseph R. Biberstine, and Katy Börner. "Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach". In: *PLoS ONE* 8.3 (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0058779.

[99] John Stasko et al. "Evaluation of space-filling information visualizations for depicting hierarchical structures". In: *International Journal of Human Computer Studies* 53.5 (2000), pp. 663–694. ISSN: 10715819. DOI: 10.1006/ijhc.2000.0420.

[100] Nick Strayer. *Physics based t-SNE*. 2018. URL: https://observablehq.com/@nstrayer/physics-based-t-sne (visited on 04/04/2019).

[101] Ito Takayuki. *Treemap in v4 - bl.ocks.org*. 2016. URL: https://bl.ocks.org/ganezasan/52fced34d2182483995f0ca3960fe228 (visited on 05/12/2019).

[102] J. B. Tenenbaum, V. De Silva, and J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. ISSN: 00368075. DOI: 10.1126/science.290.5500.2319. URL: https://science.sciencemag.org/content/290/5500/2319.full.

[103] *TensorFlow*. URL: https://www.tensorflow.org/.

[104] Anne Treisman. "Preattentive processing in vision". In: *Computer Vision, Graphics and Image Processing* 31.2 (1985), pp. 156–177. ISSN: 0734189X. DOI: 10.1016/S0734-189X(85)80004-9.

[105] Anthony Trippe. *Guidelines for Preparing Patent Landscape Reports*. Tech. rep. World Intellectual Property Organization, 2015. URL: http://www.wipo.int/tisc/en/.

[106] Eduard Trott. *Zoomable Sunburst on d3.js v4 - bl.ocks.org*. URL: https://bl.ocks.org/maybelinot/5552606564ef37b5de7e47ed2b7dc099 (visited on 05/12/2019).

[107] Thomas S. Tullis and Jacqueline N. Stetson. "A comparison of questionnaires for assessing website usability". In: *Usability Professional Association Conference* June (2004), pp. 1–12. ISSN: 0950-0782. DOI: 10.1080/09500782.2014.944427. URL: http://home.comcast.net/%7B~%7Dtomtullis/publications/UPA2004TullisStetson.pdf.

[108] C. Ware. *Information Visualization: Perception for Design*. Interactive Technologies. Elsevier Science, 2004. ISBN: 9780080478494. URL: https://books.google.de/books?id=ZmG%5C_FiqqyqgC.

[109] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. "How to Use t-SNE Effectively". In: *Distill* (2016). DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

[110] S. J. Westerman and T. Cribbin. "Mapping semantic information in virtual space: Dimensions, variance and individual differences". In: *International Journal of Human Computer Studies* 53.5 (2000), pp. 765–787. ISSN: 10715819. DOI: 10.1006/ijhc.2000.0417.

[111] Ian Wetherbee. *Google Patents Public Datasets: connecting public, paid, and private patent data | Google Cloud Blog*. 2017. URL: https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data (visited on 11/20/2018).

[112] Kent Wittenburg and Georgiy Pekhteryev. "Multi-Dimensional Comparative Visualization for Patent Landscaping". In: *IEEE VIS Workshop 2015 BusinessVis15*. Chicago, 2015.

[113] World Intellectual Property Organization (WIPO). *World Intellectual Property Indicators 2017*. 2017, pp. 60–97. ISBN: 9789280521528. DOI: 10.1016/0172-2190(79)90016-4. arXiv: 31. URL: http://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo%7B%5C_%7Dpub%7B%5C_%7D941%7B%5C_%7D2013.pdf.

[114] Jiang Zhao, Man Lan, and Jun Feng Tian. "ECNU: Using Traditional Similarity Measurements and Word Embedding for Semantic Textual Similarity Estimation". In: SemEval (2015), pp. 117–122. DOI: 10.18653/v1/s15-2021.

[115] Jian Zhao et al. "Interactive exploration of implicit and explicit relations in faceted datasets". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2080–2089. ISSN: 10772626. DOI: 10.1109/TVCG.2013.167.

A. Appendix

A.1. Semi-structured interview questionnaire

Begrüßung - 3 Min

- Sick für die Zeit bedanken
- Vorstellung meiner Arbeit: Ich untersuche, wie man die Patent-Landscaping-Aufgabe unterstützen kann. Ich werde eine Lösung zu der Datenverarbeitung, die hinter den Kulissen passiert, anbieten, aber vor allem zu der Visualisierung. Mein Fokus liegt auf der Interaktion - wie genau man das Wichtigste aus den Daten rausholen kann.
- Ziels des Gesprächs erklären: Ich möchte verstehen, wie man mit den Daten interagiert. Nicht der Benutzer wird getestet, sondern das System. Es gibt keine richtigen und falschen Antworten.
- Ich möchte gerne unser Gespräch aufnehmen - mit einem Mikrofon, aber ohne eine Kamera. Das ist dazu da, dass wir frei reden können und ich nicht die ganze Zeit nur am Notizenmachen bin. Die Aufnahme kann nach Wunsch jederzeit gestoppt werden, genauso wie das ganze Interview. Dadurch entstehen dem Teilnehmer keine Nachteile.
- Darlegen, wie die Daten benutzt werden: Ich werde anhand von der Aufnahme unser Gespräch transkribieren und danach die Aufnahme löschen. Selbstverständlich werden die Daten nicht an Dritte weitergegeben. Ich werde sie nur im Kontext der Masterarbeit benutzen und möchte gerne das, was Sie sagen, in meiner Arbeit zitieren.
- Haben Sie Fragen zur Organisation?
- Einverständniserklärung unterschreiben lassen
- Aufnahme starten

Aufwärmfragen - 5 Min

- Stellen Sie sich bitte vor.
- Was ist Ihr beruflicher Hintergrund (in welchen Branchen haben Sie gearbeitet, wie lange)?
- Wie sieht Ihr typischer Arbeitsalltag aus?

- Wie ist Ihr Verhältnis zur Technologie allgemein? Wie vertraut sind Sie mit Technologie?
- Wie lange beschäftigen Sie sich schon mit Patentenanalyse?

Hauptteil - 45 Min

- Welche Programme benutzen Sie normalerweise in der Arbeit? (zusätzlich zu STN AnaVist)
- Wie ist die Aufteilung zwischen allen Tools (Zeit, Aufwand)?
- Können Sie mir bitte von ihrem letzten erstellten Landscape erzählen?
  - In welcher Form kommt die Aufgabenstellung für ein Landscape?
  - Welche Bedürfnisse haben die Kunden, die Landscapes beantragen?
  - Wie lange dauert es, einen Patent Landscape Bericht zu erstellen?
  - Wie viele Berichte haben Sie ca. schon erstellt?
  - Wie umfangreich ist das Ergebnis?
  - Welche Grafiken (Typ, Achsen) kommen normalerweise in einem Bericht vor?
  - Wie finden Sie die passenden Suchanfragen?
    - Wie lang und komplex sind die Anfragen?
    - Wie viele Treffer ergibt die Suche?
    - Relevanz der Ergebnisse?
  - Wie aussagekräftig sind die Beschriftungen der Cluster im Themescape?
  - Gibt es mehrere Abstraktionsebenen bei der Suche? Wenn ja, wie unterscheiden sie sich? (Sublandscapes)
  - Worauf achten Sie bei der Aufgabe?
    - Was für eine Bedeutung haben die Verbindungen zwischen einzelnen Patenten (Zitierung, Patentfamilie, gemeinsamer Autor oder Assignee)?
    - Was für eine Bedeutung hat die zeitliche Entwicklung?
    - Was für eine Bedeutung hat die hierarchische Klassifikation?
    - Welchen Einfluss hat Concept Frequency?
- Gibt es etwas, was Sie persönlich am existierendem System stört?
- Wie unterscheidet sich die Patentsuche für verschiedene Themengebiete? Gab es Fälle, wo es besonders gut oder besonders schlecht funktioniert hat?
- Gibt es etwas, was Sie gelernt haben nach einigen erstellten Landscapes?
- Gibt es interessante Geschichten/Anekdoten, die sie teilen wollen?
- Haben Sie sich mit Menschen ausgetauscht, die andere Tools verwenden? Wie waren ihre Erfahrungen?

Ausblick für die Zukunft - 3 Min

- Was wäre ein perfektes System für Patent Landscaping?
- Abschluss - 3 Min
- Gibt es sonst etwas, was Sie mir mitteilen wollen?
- Haben Sie Fragen an mich?
- Bedanken und verabschieden
- Aufnahme stoppen
- Hauptgedanken notieren

A.2. SQL query for the 3D printer dataset

```
SELECT DISTINCT p.publication_number
FROM
    ‘patents-public-data.patents.publications‘ p
    LEFT JOIN UNNEST(p.cpc) AS cpc_code,
    UNNEST(p.title_localized) AS title,
    UNNEST(p.abstract_localized) AS abstract,
    UNNEST(p.ipc) AS ipc_code
WHERE
    title.language = ’en’ AND abstract.language = ’en’
    AND
    (
         REGEXP_CONTAINS(abstract.text, r’3D’) OR REGEXP_CONTAINS(title.text, r’3D’)
      OR REGEXP_CONTAINS(abstract.text, r’3-D’)  OR REGEXP_CONTAINS(title.text, r’3D’)
      OR REGEXP_CONTAINS(abstract.text, r’3-dimension’)
      OR REGEXP_CONTAINS(title.text, r’3-dimension’)
      OR REGEXP_CONTAINS(abstract.text, r’3 dimension’)
      OR REGEXP_CONTAINS(title.text, r’3 dimension’)
      OR REGEXP_CONTAINS(abstract.text, r’three dimension’)
      OR REGEXP_CONTAINS(title.text, r’three dimension’)
    )
    AND
    (
        REGEXP_CONTAINS(abstract.text, r’desktop’)
      OR REGEXP_CONTAINS(title.text, r’desktop’)
      OR REGEXP_CONTAINS(abstract.text, r’additive’)
      OR REGEXP_CONTAINS(title.text, r’additive’)
    )
    AND
    (
        REGEXP_CONTAINS(abstract.text, r’print’)
```

```
  OR REGEXP_CONTAINS(title.text, r'print')
  OR REGEXP_CONTAINS(abstract.text, r'fabricat')
  OR REGEXP_CONTAINS(title.text, r'fabricat')
  OR REGEXP_CONTAINS(abstract.text, r'manufactur')
  OR REGEXP_CONTAINS(title.text, r'manufactur')
)
AND
(
  (
       REGEXP_CONTAINS(ipc_code.code, r'B29C')
    OR REGEXP_CONTAINS(ipc_code.code, r'H01L')
    OR REGEXP_CONTAINS(ipc_code.code, r'G06F')
    OR REGEXP_CONTAINS(ipc_code.code, r'G02B')
    OR REGEXP_CONTAINS(ipc_code.code, r'B32B')
    OR REGEXP_CONTAINS(ipc_code.code, r'H05K')
    OR REGEXP_CONTAINS(ipc_code.code, r'B41J')
    OR REGEXP_CONTAINS(ipc_code.code, r'B41M')
    OR REGEXP_CONTAINS(ipc_code.code, r'G06T')
    OR REGEXP_CONTAINS(ipc_code.code, r'B44C')
    OR REGEXP_CONTAINS(ipc_code.code, r'B22F')
    OR REGEXP_CONTAINS(ipc_code.code, r'H04L')
    OR REGEXP_CONTAINS(ipc_code.code, r'G03F')
    OR REGEXP_CONTAINS(ipc_code.code, r'H04N')
    OR REGEXP_CONTAINS(ipc_code.code, r'C04B')
    OR REGEXP_CONTAINS(ipc_code.code, r'G05B')
    OR REGEXP_CONTAINS(ipc_code.code, r'G03B35')
    OR REGEXP_CONTAINS(ipc_code.code, r'A61')
  )
  OR REGEXP_CONTAINS(cpc_code.code, r'B44C')
)
AND NOT
(
      REGEXP_CONTAINS(abstract.text, r'stereoscopic')
   OR REGEXP_CONTAINS(title.text, r'stereoscopic')
   OR REGEXP_CONTAINS(abstract.text, r'oxidation product')
   OR REGEXP_CONTAINS(title.text, r'oxidation product')
   OR REGEXP_CONTAINS(abstract.text, r'streaming interactive')
   OR REGEXP_CONTAINS(title.text, r'streaming interactive')
   OR REGEXP_CONTAINS(abstract.text, r'nanoweb')
   OR REGEXP_CONTAINS(title.text, r'nanoweb')
   OR REGEXP_CONTAINS(abstract.text, r'nano web')
   OR REGEXP_CONTAINS(title.text, r'nano web')
   OR REGEXP_CONTAINS(abstract.text, r'nanofiber')
   OR REGEXP_CONTAINS(title.text, r'nanofiber')
   OR REGEXP_CONTAINS(abstract.text, r'nanofibre')
```

```
OR REGEXP_CONTAINS(title.text, r'nanofibre')
OR REGEXP_CONTAINS(abstract.text, r'nano fiber')
OR REGEXP_CONTAINS(title.text, r'nano fiber')
OR REGEXP_CONTAINS(abstract.text, r'nano fibre')
OR REGEXP_CONTAINS(title.text, r'nano fibre')
OR REGEXP_CONTAINS(abstract.text, r'nanometer fiber')
OR REGEXP_CONTAINS(title.text, r'nanometer fiber')
OR REGEXP_CONTAINS(abstract.text, r'nanometer fibre')
OR REGEXP_CONTAINS(title.text, r'nanometer fibre')
OR REGEXP_CONTAINS(abstract.text, r'non halogen')
OR REGEXP_CONTAINS(title.text, r'non halogen')
OR REGEXP_CONTAINS(abstract.text, r'non-halogen')
OR REGEXP_CONTAINS(title.text, r'non-halogen')
OR
(
   (
         REGEXP_CONTAINS(abstract.text, r'food')
      OR REGEXP_CONTAINS(title.text, r'food')
      OR REGEXP_CONTAINS(abstract.text, r'feed')
      OR REGEXP_CONTAINS(title.text, r'feed')
      OR REGEXP_CONTAINS(abstract.text, r'liquid')
      OR REGEXP_CONTAINS(title.text, r'liquid')
      OR REGEXP_CONTAINS(abstract.text, r'rheolog')
      OR REGEXP_CONTAINS(title.text, r'rheolog')
   )
   AND REGEXP_CONTAINS(abstract.text, r'additive')
   OR REGEXP_CONTAINS(title.text, r'additive')
)
OR REGEXP_CONTAINS(abstract.text, r'seed culture')
OR REGEXP_CONTAINS(title.text, r'seed culture')
OR REGEXP_CONTAINS(abstract.text, r'nanometre fiber')
OR REGEXP_CONTAINS(title.text, r'nanometre fiber')
OR REGEXP_CONTAINS(abstract.text, r'nanometre fibre')
OR REGEXP_CONTAINS(title.text, r'nanometre fibre')
OR REGEXP_CONTAINS(abstract.text, r'antibacteria')
OR REGEXP_CONTAINS(title.text, r'antibacteria')
OR REGEXP_CONTAINS(abstract.text, r'media access control')
OR REGEXP_CONTAINS(title.text, r'media access control')
OR REGEXP_CONTAINS(abstract.text, r'multi-wafer 3D CAM cell')
OR REGEXP_CONTAINS(title.text, r'multi-wafer 3D CAM cell')
OR REGEXP_CONTAINS(abstract.text, r'3-sigma')
OR REGEXP_CONTAINS(title.text, r'3-sigma')
OR REGEXP_CONTAINS(abstract.text, r'three sigma')
OR REGEXP_CONTAINS(title.text, r'three sigma')
OR REGEXP_CONTAINS(abstract.text, r'vibration isolator')
```

```
       OR REGEXP_CONTAINS(title.text, r'vibration isolator')
    )
GROUP BY p.publication_number, title.text, abstract.text;
```

A.3. SQL query for the contact lens dataset

```
SELECT DISTINCT
    REGEXP_EXTRACT(LOWER(p.publication_number), r'\w+-(\w+)-\w+') as pub_num
FROM
    `patents-public-data.patents.publications` p,
    UNNEST(p.title_localized) AS title,
    UNNEST(p.abstract_localized) AS abstract,
    UNNEST(p.ipc) AS ipc_code
WHERE
    title.language = 'en'
    AND abstract.language = 'en'
    AND (
           REGEXP_CONTAINS(abstract.text, r'contact lens')
           OR REGEXP_CONTAINS(title.text, r'contact lens')
    )
    AND (
           ipc_code.code = 'G02C7/00' OR ipc_code.code = 'G02C7/02'
        OR ipc_code.code = 'G02C7/04' OR ipc_code.code = 'G02C7/06'
        OR ipc_code.code = 'G02C7/08' OR ipc_code.code = 'G02C13/00'
    )
    AND p.country_code = 'US'
GROUP BY p.publication_number, title.text, abstract.text;
```

A.4. Plan for the summative study

1. Einführung - 10 Min.
2. Aufgabenteil 1 - 20 Min.

   Lösen Sie bitte folgende Aufgaben. Versuchen Sie dabei laut zu denken und Ihre Vorgänge zu beschreiben.

   a) Welche IPC-Klassen (auf Section-Ebene, 1 Buchstabe) treten oft zusammen auf?

   b) In welchem Zeitintervall entwickelt sich der Bereich *G02C13 (Zusammenbau, Reparatur und Reinigung der Kontaktlinsen)* aktiv?

   c) Wählen Sie einen Assignee aus den 3 größten. Kooperiert diese Einrichtung viel mit anderen? Wenn ja, sind es eher andere Einrichtungen oder Privatpersonen?

d) Vergleichen Sie den zeitlichen Verlauf der Anmeldeaktivität von *Johnson & Johnson Vision Care* und *Bausch & Lomb*.

e) Vergleichen Sie, in welchen IPC-Bereichen *Novartis AG* und *Johnson & Johnson Vision Care* aktiv sind.

f) Navigieren Sie durch Verschieben und Reinzoomen in den Datensatz rein und wieder raus und beurteilen Sie dabei die Lesbarkeit der Beschriftungen bei unterschiedlichem Detaillierungsgrad.

Füllen Sie bitte den Fragebogen zur Benutzbarkeit aus.

3. Aufgabenteil 2 - 20 Min.

   Lösen Sie bitte folgende Aufgaben. Versuchen Sie dabei laut zu denken und Ihre Vorgänge zu beschreiben.

   Ansatz 1

   a) **Für diese Teilaufgabe bitte Hierarchie oben links auf „Country" umschalten.** Finden Sie den Bereich / die Bereiche mit Patenten über farbige Kontaktlinsen

   b) Finden Sie den Bereich / die Bereiche mit Patenten über Kontaktlinsen mit elektronischen Komponenten ("Smart"e Kontaktlinsen)

   c) Beschreiben Sie kurz die groben thematischen Bereiche im Datensatz (Große Cluster) mit eigenen Worten. Beurteilen Sie die Platzierung der Bereiche zueinander.

   Füllen Sie bitte den Fragebogen zum Vergleich der Ansätze aus.

   Schalten Sie bitte nun in der linken oberen Ecke den Ansatz um – wenn bis jetzt A eingestellt wurde, dann auf B umschalten, und anders rum.

   Ansatz 2

   a) **Für diese Teilaufgabe bitte Hierarchie oben links auf „Country" umschalten.** Finden Sie den Bereich / die Bereiche mit Patenten über Reinigung der Kontaktlinsen

   b) Finden Sie den Bereich / die Bereiche mit Patenten über Bestellungssysteme für Kontaktlinsen (es geht um die Interaktion mit dem Kunden, also Diagnose, Bestellung z.B. als Abo, Anpassung des Rezepts etc.)

   c) Beschreiben Sie kurz die groben thematischen Bereiche im Datensatz (Große Cluster) mit eigenen Worten. Beurteilen Sie die Platzierung der Bereiche zueinander.

   Füllen Sie bitte den Fragebogen zum Vergleich der Ansätze aus.

4. Besprechung der Ergebnisse - 10 Min.

A.5. Comparison of extracted terms for semantic and baseline approaches

The tables are in the descending order sorted by size of the dataset. Matching terms are emphasized in bold.

Table A.1.: Comparison of cluster key terms for both approaches. Diesel engines dataset

No	Top 15 key terms per approach	
	Semantic	Baseline
1	**electrode**, **tube**, **reactor**, **module**, **conduit**, **plasma**, **exchanger**, **heat exchanger**, **work vehicle**, **burner**, **arrangement**, egr, **duct**, **cleaning**, dosing	**electrode**, **tube**, **burner**, **reactor**, **conduit**, **module**, **plasma**, **work vehicle**, **exchanger**, **heat exchanger**, **cleaning**, **duct**, **arrangement**, electrodes, bracket
2	**s_number_**, **exhaust purification**, **purification apparatus**, **forced regeneration**, **control apparatus**, **abnormality**, **mode**, **selective reduction**, reduction catalyst, **particle filter**, diagnosis, judgment, accumulation amount, injection control, **purifying system**	**s_number_**, **purification apparatus**, urea, reductant, **exhaust purification**, egr, urea water, **control apparatus**, **forced regeneration**, **abnormality**, **selective reduction**, **purifying system**, dosing, **particle filter**, mode
3	**ceramic honeycomb**, **honeycomb filter**, **mat**, **article**, **plugging**, **plugged honeycomb**, **honeycomb segment**, **honeycomb structural**, **bonding**, **segments**, **structural**, **bonding material**, **honeycomb structured**, **structured body**, porous body	**ceramic honeycomb**, **mat**, **honeycomb filter**, **plugging**, **honeycomb structural**, **plugged honeycomb**, **honeycomb segment**, **structural**, **article**, **bonding**, **segments**, pollution control, **bonding material**, **structured body**, **honeycomb structured**
4	**lubricating**, **lubricating oil**, **oil composition**, **catalyst composition**, **washcoat**, **purification catalyst**, **composite oxide**, **zeolite**, **composite**, **composition**, **purifying catalyst**, **molecular sieve**, platinum group, support material, **zone**	**lubricating**, **lubricating oil**, **oil composition**, **catalyst composition**, **washcoat**, **zone**, **composite oxide**, **zeolite**, **composite**, **composition**, **purification catalyst**, **molecular sieve**, acid, **purifying catalyst**, sieve

A.6. Figures

Table A.2.: Comparison of cluster key terms for both approaches. Video codec dataset

No	Top 15 key terms per approach	
	Semantic	Baseline
1	**intra prediction**, **split**, **chroma**, **luma**, **quantization parameter**, filtering, depth, strength value, **filter strength**, **coding units**, **prediction mode**, **prediction modes**, **mpm**, scanning, **block boundary**	**split**, **filtering**, **depth**, **strength value**, **coding units**, **filter strength**, maximum coding, **block boundary**, pixels block, **strength**, successive pixels, boundary, split information, transformation unit, transformation
2		**intra prediction**, **chroma**, merge, **candidate**, **luma**, **quantization parameter**, target, merge candidate, **prediction mode**, **prediction modes**, **mpm**, target block, samples, motion information, pair
3	**string**, **video picture**, **unequal**, **header information**, **packets**, **frames**, **code table**, **compressed**, **referenceable**, order value, event, bidirectional, length code, image data, identifier	scanning, layer, **unequal**, **string**, **header information**, **video picture**, **compressed**, packets, context model, **code table**, model, scanning pattern, **referenceable**, object, **frames**
4	**rounding, rounding information, bilinear interpolation, bilinear, prediction image, current frame, coded information, synthesizing prediction, synthesizing, encoded bitstream, bitstream current, dct coefficients, dct, frame current, values specifies**	**rounding, rounding information, bilinear interpolation, bilinear, current frame, prediction image, coded information, synthesizing prediction, synthesizing, encoded bitstream, bitstream current, dct coefficients, dct, frame current, values specifies**
5	**located block**, target, candidate, **reference frame**, vector predictor, merge, **list**, **picture index**, **neighboring blocks**, target block, **weighting**, **vector located**, **decoded picture**, merge candidate, layer	**located block**, **reference frame**, **picture index**, **vector located**, **weighting**, predictive block, picture list, list motion, frame picture, **neighboring blocks**, **list**, predictive image, vector predictive, weighting factor, **decoded picture**

Table A.3.: Comparison of cluster key terms for both approaches. 3D printer dataset

No	Top 15 key terms per approach		Our comments
	Semantic	Baseline	
1	**resin**, **ink**, **curable**, powder, **precursor**, **composition**, **reactive**, **water**, **monomer**, **slurry**, parts, particulate, article, diluent , polymerizable	**resin**, **ink**, **mold**, **curable**, **precursor**, **sheet**, **reactive**, **water**, **monomer**, **slurry**, molding, **composition**, cement, composite, ceramic	Materials for printing, especially resin
2	**filament**, **build**, mold, **nozzle**, **cooling**, **powder**, **print**, sheet, **dielectric**, component, cavity, module, **print head**, **head**, channel	**filament**, **build**, **powder**, _number_d printer, **nozzle**, **print**, **cooling**, build material, **dielectric**, desktop, **head**, **print head**, channel, interconnect, printing device	Printing process and materials for printing, especially plastic filament
3	**implant**, **bone**, **dental**, patient, **teeth**, anatomy, distal, oral, digital model, custom, mold, jaw, porous, denture, graft	**implant**, **pliim**, **pliim based**, **planar laser**, **laser illumination**, **image detection**, **plib**, **image formation**, **detection**, **supportable**, **bone**, **detection array**, **dental**, **teeth**, _number_d model	3D-printed dental prosthetics and other medical applications. PLIIM stands for Planar Laser Illumination and Imaging which is used in 3D scanning
4	**pliim**, **pliim based**, **planar laser**, **laser illumination**, **image detection**, **plib**, **image formation**, **detection**, **supportable**, **detection array**, _number_d object, virtual, communications, simulation, information		

Table A.4.: Comparison of cluster key terms for both approaches. Hair dryer dataset

No	Top 15 key terms per approach	
	Semantic	Baseline
1	**circuit**, **temperature**, **power**, hot air, appliance, **generator**, casing, **current**, **barrel**, **voltage**, **speed**, **cord**, **channel**, **airflow**, **sensor**	**circuit**, **power**, **temperature**, filter, **cord**, **speed**, battery, **sensor**, blower, light, source, system, **current**, power cord, blow dryer
2		**barrel**, **generator**, blow dryer, **channel**, ion, blow, attachment, unit, **airflow**, **voltage**, bottom, holder, portable, stand, frame
3	**assembly**, **head**, blow, attachment, apparatus, **hood**, blow dryer, bottom, **shell**, **arm**, top, barrel, plate, holder, configuration	apparatus, **assembly**, **hood**, **head**, duct, accordance, **shell**, chamber, valve, conduit, grip, **arm**, pivot, tubular, mounting
4	**fluid**, **hairdryer**, **path**, **appliance**, **duct**, emitting, sleeve, liquid, ion, components, water, fixed, combination, hollow, results	diffuser, plate, **duct**, **fluid**, hot air, shell, connector, attachment, external, **hairdryer**, **path**, **appliance**, face, sheet, plastic
5	**vanes**, **duct**, **impeller**, **casing**, diffuser, ring, **plate**, outlet opening, tubular, guide, hood, shaft, blow dryer, inlet opening, barrel	**casing**, **impeller**, wire, chamber, **vanes**, tube, longitudinal axis, longitudinally, port, guide, extension, axial, part, **duct**, **plate**

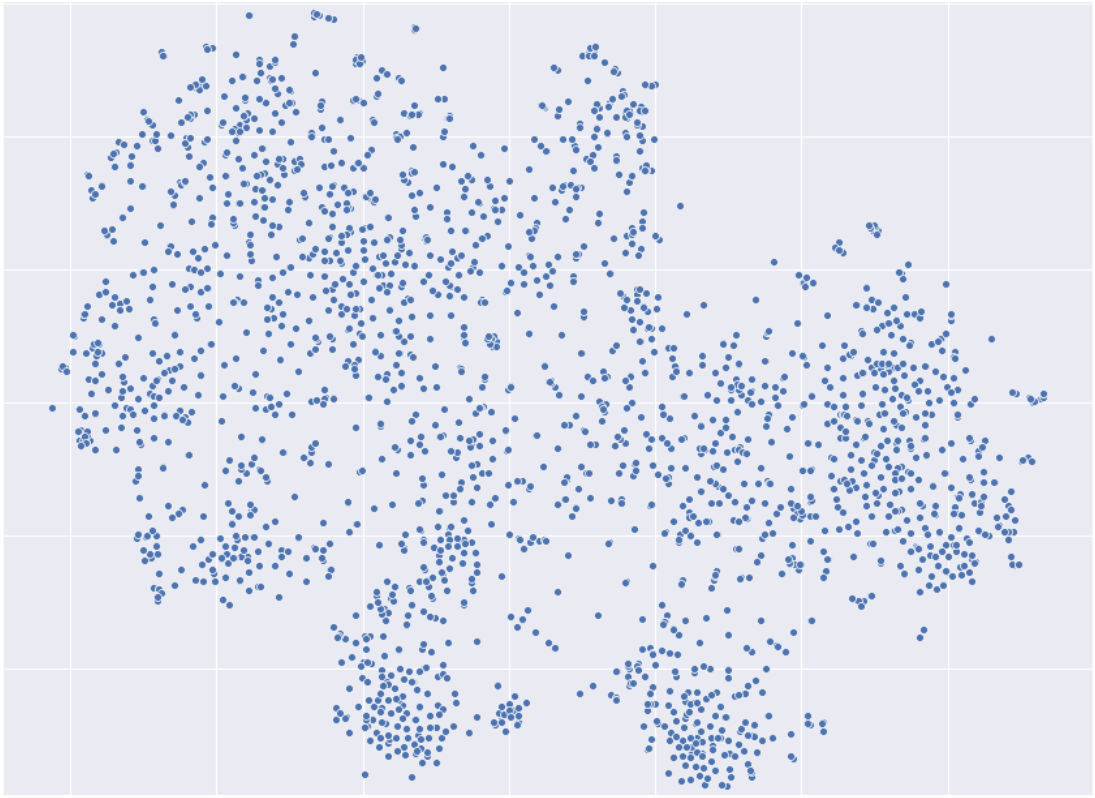
(a) Non-metric MDS

(b) Metric MDS

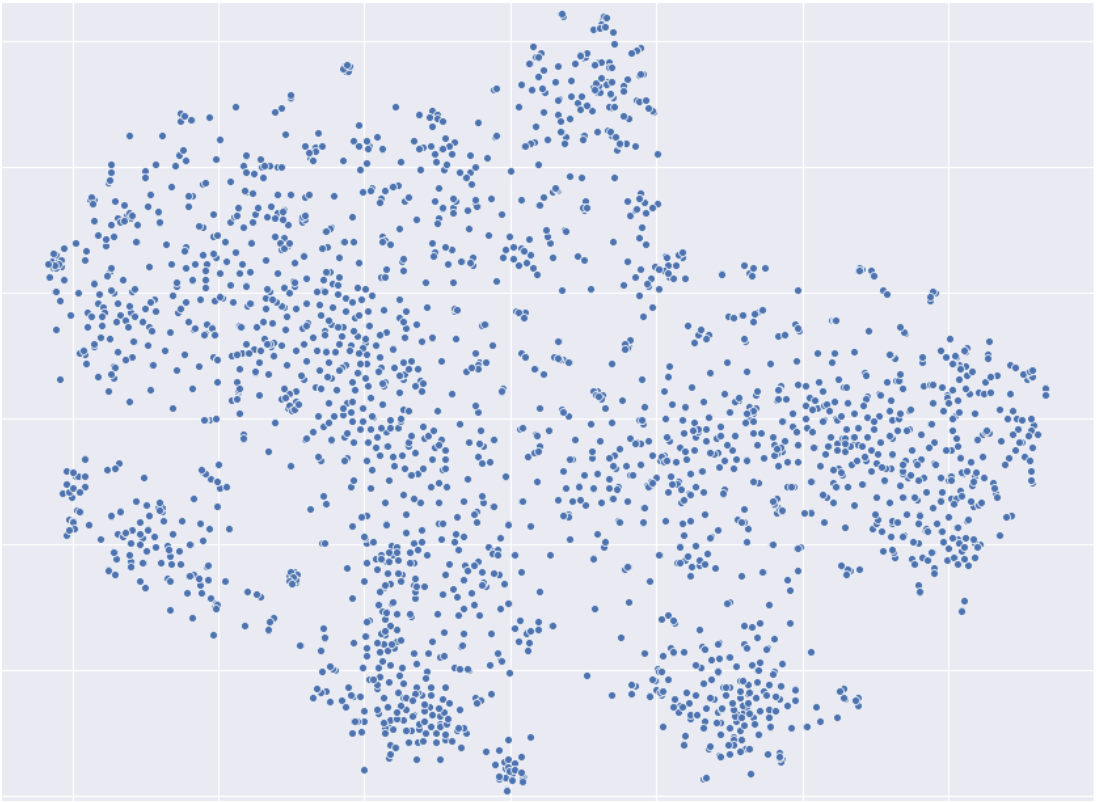
(c) Isomap. Straight lines represent patent families.

(d) PCA

Figure A.1.: A comparison of dimension reduction techniques applied to the video codec dataset.

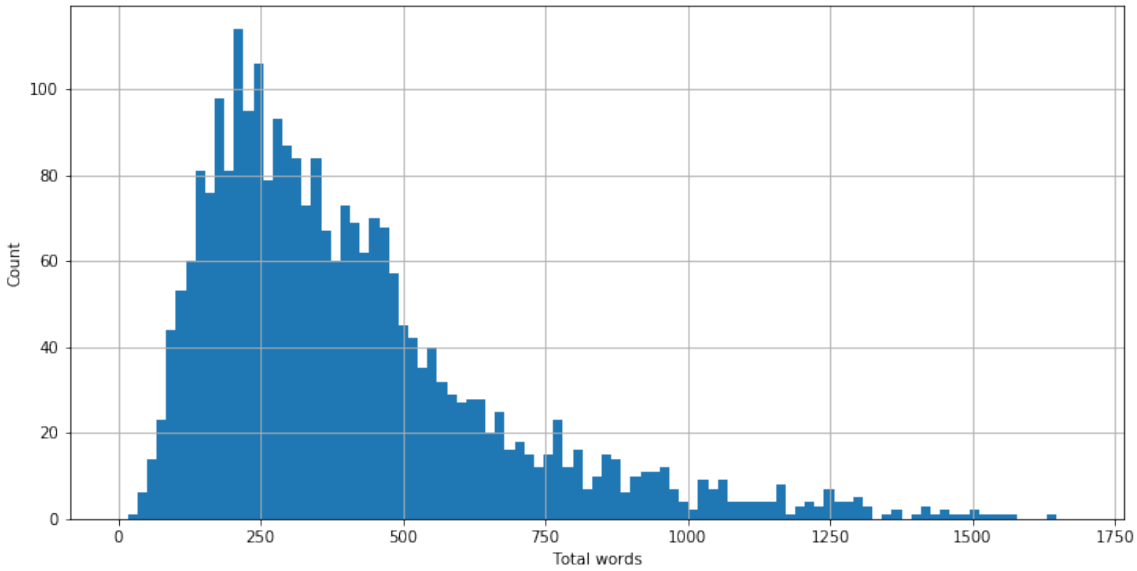


(a) Non-weighted average

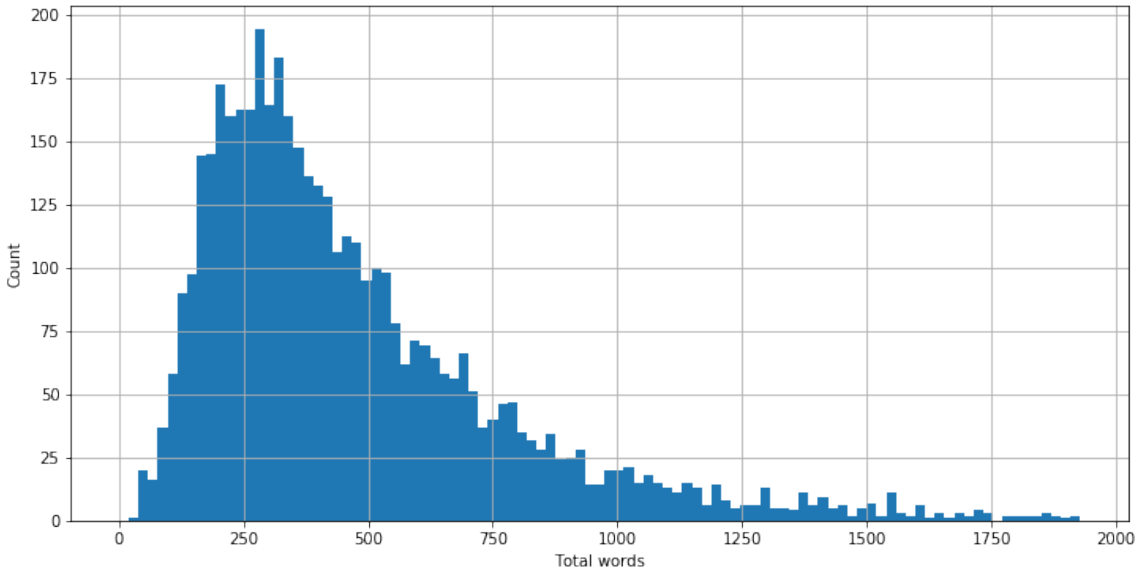


(b) Weighted average

Figure A.2.: A comparison of document vectors computed with and without IDF weighting. Contact lens dataset.



(a) Contact lens dataset



(b) Diesel dataset

Figure A.3.: Distributions of text length after stopword removal