ScholarSight: Visualizing Temporal Trends of Scientific Concepts

Michael Färber
University of Freiburg
Freiburg, Germany
michael.faerber@cs.uni-freiburg.de

Chifumi Nishioka
Kyoto University
Kyoto, Japan
nishioka.chifumi.2c@kyoto-u.ac.jp

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

ABSTRACT

In this paper, we present a system for exploring the temporal trends of scientific concepts. Scientific concepts were captured by extracting noun phrases and entities from all computer science papers of arXiv.org. Our system allows users to review the time series of numerous concepts and to identify positively and negatively trending concepts. By applying clustering techniques and cluster analysis visualizations, it can also present concepts which share the same usage patterns over time. Our system can be beneficial for both ordinary researchers of any field and for researchers working in bibliometrics and scientometrics in order to investigate the evolution of scientific concepts.

KEYWORDS

trend detection, scholarly data, bibliometrics, time series

1 MOTIVATION

The number of researchers and scientific publications worldwide in all disciplines has increased dramatically. Also, scientific concepts are subject to constant change. New scientific concepts emerge and either replace existing concepts or are related to them. We argue that this phenomenon of scientific concept evolution is not only relevant to bibliometrics and scientometrics researchers (i.e., researchers studying the evolution and behavior in science), but also to ordinary researchers, as they might be interested in obtaining answers to the following questions:

Q1: **Time Series Review:** Given a scientific concept, how often does it appear in scientific papers over time?

Q2: **Similar Usage Patterns:** Which concepts follow similar usage patterns over time? How are they characterized?

Q3: **Positive and Negative Trends:** Which scientific concepts have become commonly used in recent years, and which ones have become infrequently used?

In the following, we present our framework that addresses the above-mentioned aspects. The framework is available online at **http://scholarsight.org/** and its source code is available online at **http://github.com/michaelfaerber/scholarsight**.

2 CONCEPT EVOLUTION ANALYSIS

We now describe our approach for extracting concepts from scientific papers and identifying positive and negative trends.

Data Set. We use the *arXiv CS data set* [2] as our database. This data set contains the plain texts of all papers hosted at arXiv.org in the field of computer science. In total, the data set covers about 90,000 papers, resulting in about 16 million plain text sentences. Note that in this dataset, formulas have been replaced by corresponding placeholders for easier text processing.

We are interested in the concepts mentioned in the papers. Thus, we apply the following two concept extraction techniques:

Extracting Noun Phrases. We extract noun phrases from the papers' plain texts. Our approach uses rules on the part-of-speech tags obtained by the Stanford parser. Given the 15.5M sentences from the initial data set, we collect 10.7M unique noun phrases (76.7M non-unique).

Extracting Entity Mentions. Noun phrases are quite an intuitive way of extracting concepts from text. However, using noun phrase extraction, the problem persists that ambiguities in the language are not resolved.[1] Thus, we also automatically annotate all papers in our data set with Wikipedia URIs (e.g., linking "CNN" to Convolutional_Neural_Network[2]), using the state-of-the-art text annotation service x-LiSA [9]. Given the 15.5M sentences, we obtain 25.8M (non-unique) entity mentions, which link to 151,529 unique Wikipedia URIs.

Filtering Time Series. Processing all of the extracted noun phrases and entities results in very large databases and declined querying performance. Thus, we filter concepts as follows (following the similar procedure of [1]): (1) each concept needs to appear in at least 100 documents within the whole corpus; (2) each concept needs to appear in at least three different years.

Identifying Trending Concepts and Concept Changes. Our framework contains a component to detect positively and negatively trending concepts over time based on the Mann-Kendall test [3]. Among the most positively trending noun phrases are "regularizer," "ground truth," "GPUs," and "machine learning techniques." However, concepts such as "Wikipedia," "one-shot learning," and even "LDA" have also been used with increasing frequency. Among the most negatively trending noun phrases are "block length," "bits," "Shannon," and "message." Using the list of positively trending concepts and the list of negatively trending concepts, changes and replacements of concepts over time can be identified. For instance,

[1] For instance, phrases with different senses are grouped together (e.g., "CNN" referring to the TV station and to the convolutional neural network), while different phrases referring to the same concepts are treated independently (e.g., "convolutional neural network," "CNN," etc.).

[2] We omit the prefix http://en.wikipedia.org/wiki/ when referring to Wikipedia URIs.

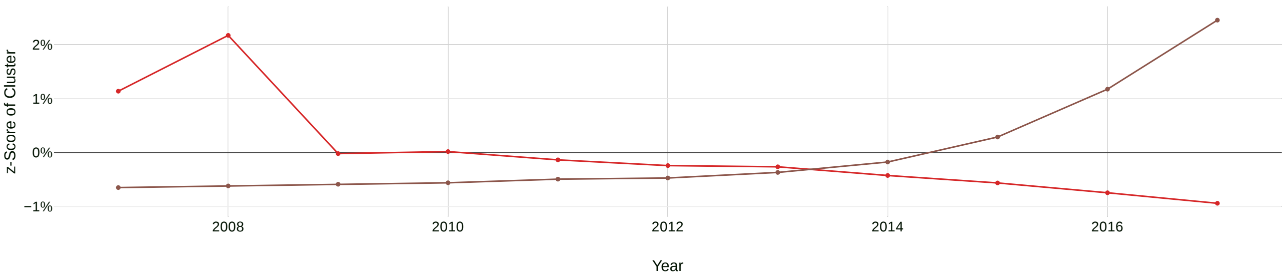

Figure 1: Time series of the centroid of cluster 13 (decreasing at the end; containing "association rules") and cluster 35 (increasing over the time; containing "recurrent neural network" and "convolutional neural network").

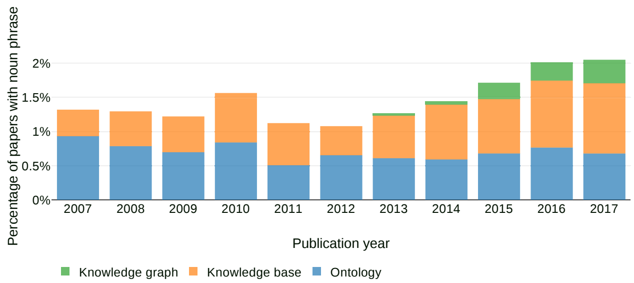

Figure 2: Relative frequency (in %) of documents containing the noun phrases "ontology," "knowledge base," and "knowledge graph," per year.

"association rules" are taken over by "neural network" and "transfer learning."

Concept Usage Pattern Analysis. We are also interested in identifying common patterns of concept usage over time (clustering various concepts together) and knowing which usage pattern (modeled as a cluster) any given concept belongs to. To this end, we employ clustering over the time series of noun phrases. Initially, we considered employing k-Shape [4], since it is a state-of-the-art clustering method using shape similarity. However, similarly shaped time series that are time-delayed still obtain a relatively high similarity. Thus, this method is not appropriate for our purpose, which is to identify which concepts evolve together. Ultimately, we normalize the time series data by z-scores, as done in [4], and apply k-means clustering with $k = 50$. This gives us relatively comprehensible clusters. For instance, we obtain a cluster with ascending trend that includes the concepts “neural network,” “recursive neural network,” and “convolutional network.” Furthermore, another cluster with descending trend contains concepts (e.g., “association rules”) that were substituted by other machine learning methods.

3 VISUALIZATION

Fig. 2 shows a snapshot of the user interface when searching for noun phrases. We realize that searching with noun phrases in some cases results in similar visualizations (i.e., same usage patterns). However, searching via Wikipedia concepts allows us to resolve ambiguities in the language. Fig. 1 shows the trends of two cluster centroids generated from the noun phrase time series data.

4 RELATED WORK

Various papers presenting approaches and demonstration systems deal with the evolution of research topics over time [6–8]. While these works primarily consider very generic research topics, we also cover very specific concepts. Furthermore, on many occasions the authors apply methods based on community networks etc. [7] rather than operating purely on the text.

In the past, several kinds of information extraction techniques have been applied to scientific papers, ranging from noun phrase extraction over entity annotation to relation extraction (see, for instance, the SemEval 2010 Task 5 and the SemEval 2017 Task 10). However, no paper dedicated to the analysis of extracted noun phrases and entities that would describe a working system has been presented to our knowledge.

5 CONCLUSION

In this paper, we have presented a framework for reviewing scientific concepts concerning their appearance over time. Based on statistics, we identified positively and negatively trending scientific concepts and showed these with the temporal course to the user. We have also considered the usage patterns of concepts over time. For the future, we plan to incorporate other paper corpora [5] and to automatically find surprising patterns in the time series data.

REFERENCES

[1] Tal Daniel and Mark Last. 2016. Exploring Long-Term Temporal Trends in the Use of Multiword Expressions. In *Proceedings of the 12th Workshop on Multiword Expressions (MWE@ACL’16)*.

[2] Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. A High-Quality Gold Standard for Citation-based Tasks. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC’18)*.

[3] Michael Färber and Adam Jatowt. 2019. Finding Temporal Trends of Scientific Concepts. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR’19)*.

[4] John Paparrizos and Luis Gravano. 2015. k-shape: Efficient and accurate clustering of time series. In *SIGMOD*. ACM, 1855–1870.

[5] Tarek Saier and Michael Färber. 2019. Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR’19)*.

[6] Angelo Antonio Salatino, Francesco Osborne, and Enrico Motta. 2018. AUGUR: Forecasting the Emergence of New Research Topics. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL’18)*. 303–312.

[7] Xiaoguang Wang, Qikai Cheng, and Wei Lu. 2014. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. *Scientometrics* 101, 2 (2014), 1253–1271.

[8] Changhong Zhang, Zeyu Li, and Jiawan Zhang. 2018. A survey on visualization for scientific literature topics. *J. Visualization* 21, 2 (2018), 321–335.

[9] Lei Zhang and Achim Rettinger. 2014. X-LiSA: Cross-lingual Semantic Annotation. *PVLDB* 7, 13 (2014), 1693–1696.