Finding Temporal Trends of Scientific Concepts

Michael Färber[1] and Adam Jatowt[2]

[1]Department of Computer Science, University of Freiburg, Germany
michael.faerber@cs.uni-freiburg.de
[2]Department of Social Informatics, Kyoto University, Japan
adam@kuis.db.kyoto-u.ac.jp

Abstract. Science evolves very rapidly, and researchers have studied the evolution of coarse-grained research topics. However, to our knowledge, no analysis of the temporal trends of fine-grained scientific concepts has been performed based on papers' full texts. For this paper, we extract noun phrases as concepts from all computer science papers of arXiv.org. We then identify positive and negative trends by means of simple linear regression, Mann-Kendall test, and Theil-Sen estimate. In our experiments, we obtain noteworthy findings about trends using the Mann-Kendall test, while the Theil-Sen estimate and simple linear regression lead to many non-scientific concepts. Our findings are potentially relevant for both ordinary researchers and researchers working in bibliometrics and scientometrics.

Keywords: trend detection, scholarly data, bibliometrics, time series

1 Motivation

Science evolves very rapidly and the increasing number of researchers and scientific publications worldwide in various disciplines reinforce this effect [1,2,3]. We argue that this phenomenon of scientific evolution is worth investigating in more detail [4,5,6,7,8,9,10]. Specifically, ordinary researchers, as well as researchers working in bibliometrics and scientometrics, might be interested in the answers to the following questions:

Q1: **Positive and negative trends:** Which scientific concepts have become common in recent years and which scientific concepts have become less common? Which concepts have maintained their relevance over time?
Q2: **Replacement of concepts:** Which concepts have recently been supplanted by other concepts?

In this work, we target those questions by extracting noun phrases from a corpus of scientific papers, namely the contents of all computer science papers of arXiv.org [11]. Then, positive and negative trends over time in the set of extracted noun phrases are identified, contributing to answering Q1. Furthermore, concepts that have replaced other concepts over time (reflected in the usage statistics) are identified, contributing to answering Q2.

2 Trend Detection

To find positive and negative trends in time series data, a variety of algorithms are available [12,13]. In the following, we focus on those we used in our analysis (see Sec. 3).

We consider years as intervals of the time series. Furthermore, we use the normalized relative frequency of the concepts as the basis for our calculations. Formally, let $D = \{d_1, ..., d_{|D|}\}$ be our document corpus. Let c_i be a concept in the concept set C occurring n_{c_i} times in the corpus D. Let D_{c_i} be the set of documents in which c_i occurs at least once. Then, the normalized relative frequency of c_i on the document level is defined as the ratio of documents containing c_i with respect to all documents: $rf_{c_i} = |D_{c_i}|/|D|$.

Simple Linear Regression. For this basic trend detection method, we calculate for a given concept c_i the difference between the relative frequencies $rf_{c_i,k}$, $rf_{c_i,l}$ of two time periods k, l (e.g., year 2007 and 2017): $d = rf_{c_i,k} - rf_{c_i,l}$.

Mann-Kendall τ. To obtain statistically significant trends in time series data, the *Mann-Kendall* test [14,13] is commonly used. This test can be applied as a non-parametric test for monotonic trends. The Mann-Kendall statistic can be used as indication whether a trend exists statistically and whether it is positive or negative. More formally, the null hypothesis of Kendall's τ is that there is no trend ($H_0 : \tau = 0$). The alternative hypothesis is that there is a trend ($H_1 : \tau \neq 0$). Given that we have the concepts in a temporal order, let G_i be the number of data points after y_i that are greater than y_i. Let L_i be the number of data points after y_i that are smaller than y_i. Then, the Kendall's τ coefficient is calculated as

$$\tau = 2S/n(n-1)$$

and S is thereby defined as

$$S = \sum_{i=1}^{n-1} (G_i - L_i)$$

and corresponds to the the sum of the differences between G_i and L_i along the time series. Since we are dealing with a sufficiently large number of time slots n, we can assume normal distribution for the test statistic z [13,15] and write

$$z = \frac{\tau}{\sqrt{2(2n+5)/9n(n-1)}}$$

Theil-Sen Trend Line. The *Theil-Sen estimate* [16] can be used to estimate the slope of a trend. It can be considered a non-parametric alternative to the parametric ordinary least squares regression line. A Theil-Sen line models how the median value changes linearly with time [13]. Formally, let

$$B_n = \{\frac{y_j - y_i}{x_j - x_i} : \ x_i \neq x_j, 1 \leq i < j \leq n\}$$

The Theil-Sen estimator $\hat{\beta}_n$ is then defined as the median of all slopes in B_n, i.e., $\hat{\beta}_n = med(B_n)$ with med standing for the median.

3 Trend Detection of Scientific Concepts

We now describe our approach for extracting concepts from scientific papers and identifying positive and negative trends.[1]

3.1 Data Set and Extracting Concepts

We use the *arXiv CS data set* [11] as our database. This data set contains the plaintexts of all papers hosted at arXiv.org in the field of computer science from the early beginnings of arXiv.org until December 2017. As corpora covering the contents of research papers are rare, and the usage of arXiv.org has become increasingly common in recent years, we believe that this data set is a valid basis for concept evolution analyses. In total, the data set covers about 90,000 papers, resulting in about 16 million plaintext sentences. Note that in this data set, formulas have been replaced by placeholders for easier text processing.

Given the papers' fulltexts, we are interested in the concepts mentioned in these papers. For this paper, we use case-insensitive noun phrases as concept representations. Thus, we extract noun phrases from the papers' fulltexts. Our approach uses an extended rule set of [18] (in total, eight rules) on the part-of-speech tags assigned by the Stanford parser.

Given the 15.5M sentences from the initial data set, we obtained 10.67M unique noun phrases (76.7M non-unique noun phrases). When ordering the extracted noun phrases by absolute frequency, we can observe that domain-specific concepts, which are in the focus of this paper, occur particularly in the mid range, while functional words and phrases common for writing papers (e.g., "number," "section," "figure") appear at the top.[2] We use this observation to filter out irrelevant concepts during trend detection.

3.2 Sparsity and Thresholds

The set of extracted noun phrases still contains many irrelevant, non-scientific noun phrases. Processing all of them would result in large databases, unnecessary trend calculations, and declined querying performance of indices. Thus, we analyzed the effectiveness of several filtering methods (following the similar procedure of [15]): (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; (3) the combination of methods 1 and 2. Table 1 shows the results. We can observe that using threshold 1 (i.e., each concept must appear in at least 100 documents) allows a considerable decrease in the number of concepts. However, threshold 2 (i.e., the number of years in which each concept needs to appear) also seems to be very effective. Ultimately, we followed [15] and used the combination of (1) and (2).

[1] See `https://github.com/michaelfaerber/scholarly-trends` for our source code and [17] for a demonstration system based on our trend detection approach.

[2] The data set of extracted noun phrases is available at `https://github.com/michaelfaerber/scholarly-trends`.

Table 1: Frequencies of noun phrases when applying various thresholds.

Threshold	Frequency	Percent of All Noun Phrases
< 100 occurrences	9,769,907	99.51%
≥ 100 occurrences	47,871	0.49%
occur in < 3 years	8,945,295	91.11%
occur in ≥ 3 years	872,600	8.89%
> 100 occurrences and occur in ≥ 3 years	47,759	0.49%
> 100 occurrences and occur in < 3 years	112	0.00014%

Table 2: Top 15 positively trending noun phrases by simple linear regression (2007-2017).

Noun phrase	# Docs
experiments	26150
dataset	15111
training	13708
table	36883
methods	29840
performance	39505
data	37760
features	17831
accuracy	16638
datasets	10002
models	23789
images	12065
model	40009
training data	8444
method	38301

Table 3: Top 15 negativley trending noun phrases by simple linear regression (2007-2017).

Noun phrase	# Docs
proof	31854
theorem	31378
definition	34978
fact	53356
course	15154
lemma	22527
elements	26122
condition	20904
case	66931
whose	39785
us	52096
length	26846
notation	21878
construction	18790
sense	24674

3.3 Applying Trend Detection Methods

Given the set of filtered noun phrase series, we apply the trend detection algorithms outlined in Sec. 2, namely the simple linear regression, the Mann-Kendall test, and the Theil-Sen estimate. We thereby obtained the following findings:

Simple Linear Regression. We list the top 15 positively and negatively trending noun phrases using the simple linear regression in Table 2 and 3.[3] Very general concepts (e.g., "experiments," "dataset," and "training") show a strong increase in the usage over time in our data set. This might be surprising, but can be partially explained by the fact that our considered concepts are from 2007 and 2017; rather general concepts remain over such a long time span. Given the negatively trending concepts, it becomes apparent that concepts concerning formalism and theories were much more important in 2007 than in 2017. Overall, we can state that the simple linear regression leads partially to already relevant

[3] The full list is available online at `https://github.com/michaelfaerber/scholarly-trends`.

Table 4: Top 15 positively trending noun phrases by Mann-Kendall test.

Noun phrase	Mann-Kendall z	# Docs
training data	4.20	8444
high accuracy	4.20	2182
regularizer	4.20	1257
pixels	4.20	5817
supplementary material	4.20	1409
liu	4.20	1028
ground truth	4.20	4349
synthetic images	4.20	278
gpus	4.20	1389
hours	4.20	2277
cloud	4.20	1555
gradient	4.20	6963
higher accuracy	4.20	1206
millions	4.20	2963
machine learning techniques	4.20	776

Table 5: Top 15 negatively trending noun phrases by Mann-Kendall test.

Noun phrase	Mann-Kendall z	# Docs
block length	-4.20	937
bits	-4.20	9297
transmitted codeword	-4.20	409
course	-4.20	15154
capacity	-4.20	8716
spin glasses	-4.20	109
shannon	-4.20	1912
bit	-4.20	7945
message	-4.20	8146
codes	-4.20	6015
real numbers	-4.20	2940
cdma systems	-4.20	94
alphabet size	-4.20	570
intermediate nodes	-4.05	725
codeword	-4.05	3347

findings about trends of concepts, although many abstract concepts are also found to be trending.

Mann-Kendall test. Table 4 and Table 5 list the top 15 positively and negatively trending noun phrases using the Mann-Kendall test (see Sec. 2). Out of all 47,759 indexed noun phrases, 19,525 of them are found to have a statistically significant (positive or negative) trend over the years (using Kendall's $\tau|z| > 3$ as in [15]). This value might appear high. However, note that we have applied a strong filter for obviously irrelevant concepts (see Sec. 3.2).

Considering all trending noun phrases, we can recognize that the Mann-Kendall test appears to be a reasonable trend detection method for our case. We obtained noteworthy findings concerning the trending concepts:

- Among the positively trending concepts are many machine learning-associated concepts, such as "gradient," "deep neural networks," "convolutional neural networks," "convolutional layer," and "gpus." The metrics "ROC" and "AUC" (capitalized for better readability) are also trending.
- "One-shot learning" and "data science" are identified as positively trending and render the general orientation of computer science research in recent years.
- Negatively trending noun phrases are particularly from the area of formal (i.e., theoretical) computer science, such as the area of information theory. Representative, negatively trending concepts are "block length," "bits," "shannon," and "message," but also "decision problem" and "Turing machine." It is quite obvious that arXiv was predominated by theoretical computer science, while nowadays machine learning is the predominant field. Ultimately, this means that our database is, to some extent, unbalanced. However, we believe that it is acceptable, as it reflects the general orientation of computer science research overall over the years.

Table 6: Top 15 positive trending noun phrases by Theil-Sen slope.

Noun Phrase	Theil-Sen slope	Total # Docs.
experiments	2.55	26150
performance	2.50	39505
table	2.44	36883
dataset	2.24	15111
data	2.12	37760
methods	2.04	29840
training	1.95	13708
features	1.89	17831
accuracy	1.70	16638
parameters	1.61	33835
datasets	1.56	10002
method	1.52	38301
experiment	1.45	14829
work	1.41	55240
images	1.41	12065

Table 7: Top 15 negative trending noun phrases according by Theil-Sen slope.

Noun Phrase	Theil-Sen slope	# Docs.
theorem	-2.55	31378
proof	-2.32	31854
definition	-1.76	34978
lemma	-1.65	22527
fact	-1.40	53356
course	-1.38	15154
notion	-1.33	18398
case	-1.24	66931
length	-1.21	26846
elements	-1.18	26122
us	-1.18	52096
following theorem	-1.16	14283
construction	-1.16	18790
sense	-1.13	24674
condition	-1.09	20904

– Also, the concepts "DBScan" and "LDA" have been used with increasing frequency (statistically proven) and have remained on a stable level in recent years. This may appear surprising, as those concepts are believed to have been established for a long time and therefore might be expected to decrease.
– "Quantum computing" and "PageRank" have not been identified as trending but show a strong increase and then a plateau when being visualized over time. These concepts have a very volatile time series.
– The programming language "Scala" was on the rise and then became stable, while "Python" is still increasing in recent years.

Theil-Sen Estimate. Table 6 and Table 7, respectively, list the top positive and negative trending noun phrases according to the Theil-Sen's estimate (see Sec. 2 for its definition). We can observe that using Theil-Sen leads to many very generic concepts in the lists of trending concepts, such as "experiments" and "dataset." Thus, this trend detection method can be used to generate an upper ontology instead of showing trends of specific scientific concepts.

4 Related Work

Trend Detection Based on Scientific Papers. Various papers presenting approaches and demonstration systems deal with the evolution of research topics over time [4,5,6,7,8,9,10]. Apart from the visualization frameworks for paper collections (e.g., via maps or hierarchical views) [4,5], the approach-describing papers differ from our paper as follows: (1) the authors cluster topics and, thus, rather consider high-level concepts [5,6,9]; (2) they do not apply content-based methods, but methods based on graphs and networks, such as the citation information [10,9,8] and the author information [7]; (3) they use purely the papers' titles or abstracts but no papers' full texts [5,7,8], which makes it hard to cover also long-tail concepts.

Information Extraction from Scientific Papers. In the past, several kinds of information extraction techniques have been applied to scientific papers, ranging from noun phrase extraction over entity linking to relation extraction. Noteworthy in this context are also the SemEval tasks based on scientific papers as data sets (see SemEval 2010 Task 5 'Automatic Keyphrase Extraction from Scientific Articles" [19] and SemEval 2017 Task 10: "ScienceIE – Extracting Keyphrases and Relations from Scientific Publications" [20]). While the extraction of words and bigrams has already been applied to papers [7,8], no paper dedicated to the analysis of scientific phrases in the papers' full texts has been presented to our knowledge.

Time-Series Analysis and Trend Detection. Among the most frequently used methods for trend detection are the Mann-Kendall test and Sen's slope [13]. Related to our work is the analysis of Daniel et al. [15] concerning trending multi-word expressions in the Google Books data set. However, the domain of books differs from our domain-specific use case. Furthermore, multi-word expressions cover not only noun phrases, but also proverbs, greetings, etc.

5 Conclusion

In this paper, we have presented an analysis concerning positively and negatively trending scientific concepts. We identified statistically trending concepts included in all computer science papers of arXiv.org based on several trend detection methods. We thereby found that the Mann-Kendall test performs well for this task, while the simple regression and Theil-Sen estimate have deficits, such as detecting rather general, non-scientific concepts. Based on the trending concepts, we not only found that arXiv.org has a strong orientation towards machine learning and deep learning, but we also identified rather surprising usage patterns.

For the future, we plan to consider various scientific disciplines based on the new arXiv data set presented in [21]. Moreover, we plan to perform a deeper linguistic analysis of the arXiv papers' content. For instance, extracting, storing, and testing scientific hypotheses [22] might be a worthy task.

References

1. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11) (2015) 2215–2222
2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379) (2018) eaao0185
3. Ware, M., Mabe, M.: The STM Report: An overview of scientific and scholarly journal publishing. (2015)
4. Zhang, C., Li, Z., Zhang, J.: A survey on visualization for scientific literature topics. J. Visualization **21**(2) (2018) 321–335

5. Wang, X., Cheng, Q., Lu, W.: Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics **101**(2) (2014) 1253–1271
6. Salatino, A.A., Osborne, F., Motta, E.: AUGUR: Forecasting the Emergence of New Research Topics. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. JCDL'18 (2018) 303–312
7. Bolelli, L., Ertekin, S., Giles, C.L.: Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In: Proceedings of the 31th European Conference on Information Retrieval. (2009) 776–780
8. Jo, Y., Lagoze, C., Giles, C.L.: Detecting Research Topics via the Correlation between Graphs and Texts. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'07 (2007) 370–379
9. Small, H., Boyack, K.W., Klavans, R.: Identifying emerging topics in science and technology. Research Policy **43**(8) (2014) 1450–1467
10. Popescul, A., Flake, G.W., Lawrence, S., Ungar, L.H., Giles, C.L.: Clustering and Identifying Temporal Trends in Document Databases. In: Proceedings of IEEE Advances in Digital Libraries 2000. ADL'00 (2000) 173–182
11. Färber, M., Thiemann, A., Jatowt, A.: A High-Quality Gold Standard for Citation-based Tasks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC'18 (2018)
12. Gray, K.L.: Comparison of Trend Detection Methods. PhD thesis, University of Montana, Department of Mathematical Sciences, Missoula, MT, USA (2007)
13. Interstate Technology and Regulatory Council: Groundwater Statistics and Monitoring Compliance. Statistical Tools for the Project Life Cycle. (2013)
14. Gilbert, R.O.: Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons (1987)
15. Daniel, T., Last, M.: Exploring Long-Term Temporal Trends in the Use of Multiword Expressions. In: Proceedings of the 12th Workshop on Multiword Expressions. MWE@ACL'16 (2016)
16. Sen, P.K.: Estimates of the regression coefficient based on Kendall's tau. Journal of the American statistical association **63**(324) (1968) 1379–1389
17. Färber, M., Nishioka, C., Jatowt, A.: ScholarSight: Visualizing Temporal Trends of Scientific Concepts. In: Proceedings of the 19th ACM/IEEE on Joint Conference on Digital Libraries. JCDL'19 (2019)
18. Zhao, S., Li, C., Ma, S., Ma, T., Ma, D.: Combining POS Tagging, Lucene Search and Similarity Metrics for Entity Linking. In: Proceedings of the 14th International Conference on Web Information Systems Engineering. WISE'13 (2013) 503–509
19. Kim, S.N., Medelyan, O., Kan, M., Baldwin, T.: SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval@ACL'10 (2010) 21–26
20. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 Task 10: ScienceIE – Extracting Keyphrases and Relations from Scientific Publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation. SemEval@ACL'17 (2017) 546–555
21. Saier, T., Färber, M.: 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 (2019)
22. Baker, N.C., Hemminger, B.M.: Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. Journal of Biomedical Informatics **43**(4) (2010) 510–519