Identifying Promising Research Topics in Computer Science

Rajmund Klemiński and Przemysław Kazienko

Department of Computational Intelligence, ENGINE - The European Centre for Data Science, Faculty of Computer Science and Management,
Wrocław University of Science and Technology, Poland
{rajmund.kleminski, kazienko}@pwr.edu.pl

Abstract. In this paper, we investigate an interpretable definition of promising research topics, complimented with a predictive model. Two methods of topic identification were employed: bag of words and the LDA model, with reflection on their applicability and usefulness in the task of retrieving topics on a set of publication titles. Next, different criteria for promising topic were analyzed with respect to their usefulness and shortcomings. For verification purposes, the DBLP data set, an online open reference of computer science publications, is used. The presented results reveal potential of the proposed method for identification of promising research topics.

Keywords: research prediction, promising topic, topic modelling, DBLP

1 Introduction

The study of science itself, its trends and underlying phenomena is a complex topic. Researchers have engaged in efforts to understand the evolution of scientific research areas, emergence of topics and the structure of scientific research in both newly emerging and well-established areas. Perhaps the most exciting is the prospect of being able to accurately predict the rise and fall of individual research topics and areas. Such a look into the future is of great value to science and industry alike, plotting the path for future research activities. There has been many attempts at unveiling these future trends, rising in complexity of the solutions over the years. We attempt to take a step back and provide a process and results that are easy to understand, interpret and take advantage of.

2 Related work

Prediction and identification of new, emerging or otherwise important topics has been researched for a while now. Interestingly, while 'hot topic' as a term appears in many scientific articles, particularly the titles, it is more often than not self-proclaimed by the authors. Quantifiable research in this area, on the other hand, commonly focuses on online communities. This includes both detection

[7] and propagation models [11]. There is a variety of approaches in the field of emerging topic detection. Scientists have used core documents [2], bibliometric characteristics [8] and Wikipedia sources [10] to identify and label new, emerging topics of research in different areas of science. Recently, complex approaches for predicting the trends and changes of the trends for topics have been developed with the use of ensemble forecasting [5] and rhetorical framing [9]. Evolution of the topics themselves has been tracked through metrics such as citations [3].

3 Data sources

There is a number of possible sources of bibliographic data available for research, varying in scientific areas covered, the selection of sources indexed as well as the terms and conditions applying to accessing the full database. Considering that DBLP is the only publicly available data set from among the sources relevant for us (Table 1), we have decided to use it in our research. It was our reasoning that, as the database covers the areas of science related to computer science, it would allow for reviewing the results personally, without the need to engage outside-field experts. The DBLP database contains a variety of bibliographic information on publications, including the list of authors, a title, publication year and journal/conference of publication. Additional details are available for select positions. Citation information, however, is missing.

Table 1: Available bibliographic data sources

Name	Areas	Availability	Coverage	Temporal coverage
Scopus	All science	Licensed	Best journals and conferences	1995 - present
Web of Science	All science	Licensed	Best journals and conferences	1900 - present
PubMed	Biology and Medicine	Licensed	Best journals and conferences	1966 - present
APS	Physics	By request	Best journals and conferences	1893 - present
Google Scholar	All science	Scrapping	All web sources	Varies by field
DBLP	Computer science	Public	All web sources	1936 - present

4 Promising topic

The notion of a promising research topic is well-known to every scientist; during our work, we encounter problems that appear "promising" to us. Such a feeling is usually driven by the experience in the field or based on the intuitive understanding of the problem in question. This makes the notion of a promising topic

hard to translate into objective measures without which, in turn, it is impossible to conduct research on the problem of predicting promising topics. We have attempted to define a promising topic in two ways, using only the basic metrics available. While we recognize that such a nebulous term has more nuance, the research presented in this paper was meant to investigate the effectiveness of a simple approach, providing a reference point for the future work.

1. **Significance** A research topic is *significant* if there are many articles being written on it in a year, or a sizable community of scientists are involved in writing such papers. In this view, the more promising a topic, the higher value of the metric capturing these features in the following period.
2. **Growth** While the significance of a topic gives us knowledge of how widespread it is, there is an argument that can be made against it. Once a topic becomes the central focus of a sizable community of scientists, it usually retains that status for years to come. In that sense, the notion of significance is a stagnant one, favouring established areas of research over those that experience their rapid development. A metric focusing on the *percentage growth* of either the number of articles published or the quantity of community involved in the research should capture the nuances more accurately.

5 Topics

1. **Topic identification** The first step on the road to prediction of promising topics is to identify research topics present within the database of articles. Following that, each article in the database can be described by a subset of identified topics, denoting the research areas said paper pertains to. Similarly to our approach in the case of defining a promising topic, we have selected two methods of identifying topics, differing in complexity.
2. **Bag of words.** The initial approach was to use a BoW model on the corpus of article titles obtained from the DBLP database. Considering the low value of uni-grams in the task of identifying research topics, we elected to only consider bi-grams. This is further rationalized by the fact that typically, a topic can be captured in a one- or two-word description.
3. **Latent Dirichlet allocation** – LDA, a widely used statistical model designed to identify topics within the corpus, was the second method we have employed [1]. It is a significantly more complex solution compared to bag of words, introducing its own advantages and problems. In particular, as shown in [4], the LDA model suffers from a sparsity problem when used on a corpus of short texts. To combat this problem, we have modified our corpus according to our idea of n-gram corpus, further described in 3.
4. **Topic features** In order to attempt our predictions, we have calculated several features describing the topics identified within our corpus. Note that these are based on the information available in the DBLP data set which is relatively sparse when it comes to quantifiable data.

1. **Document frequency for topic (DFT)** Denotes the number of documents within the data set that were identified as a part of a given topic. This feature is calculated separately for every year considered in our research.
2. **Δ DFT** A relative change between the value of DFT between two years, calculated for each pair of consecutive years.
3. **Distinct author frequency (DAF)** Denotes the number of unique authors present in the set of all documents assigned to a given topic. This feature is calculated separately for every year.
4. **Δ DAF** Analogously to the Δ DFT, this feature expresses the relative change in DAF. Calculated for every pair of consecutive years.
5. **Term frequency - inverse document frequency (TF-IDF)** We have employed the popular in bag of words measure of TF-IDF for both of the models; it is easily achievable for the topics identified by the LDA model by a simple analogy. This feature is calculated separately for every year.
6. **Popularity** A feature describing how many documents pertained to the given topic, relative to all documents published during a given year. Calculated for every year separately.

6 Research topic prediction

Fig. 1 showcases the prediction method applied in our research. Each step is further described in this section.

Fig. 1: Predictive process for promising topic identification

6.1 Corpus processing

1. **Corpus preparation.** We begin the preparation of data by creating a corpus of document titles published between the specified years Y_1 and Y_n . This will serve as the basis for further processing.
2. **Lemmatization.** In order to generalize article titles, we have decided to use a WordNet-based lemmatizer available for Python in [6]. Lemmatization was followed by filtering out of stop-words.

Fig. 2: Construction of the n-gram corpus

3. **N-gram corpus.** As mentioned previously, LDA model suffers from text sparsity if used on a collection of short texts. To combat this issue, we have constructed a special N-gram driven corpus. The idea behind this operation is simple. First, we acquire bi-grams from the document title corpus as a part of employing the bag of words model. Then, for each of the bi-grams found, we concatenate all document titles containing this bi-gram into a single position within the newly created n-gram corpus. This idea is visualized on Fig. 2. We have noticed a substantial improvement in both interpretability and cohesiveness of topics identified by LDA ran on n-gram corpus.
4. **Topic assignment.** Once the topic model has been trained on the corpus, we assign every probable topic to every document in the data set, allowing for later calculations. Note that this step is only carried out for LDA model.

6.2 Prediction

1. **Significance** Promising topic prediction in the sense of significance is achieved through the use of a regression model to predict either the DFT or DAF feature of topics, then rank them based on these features. To verify prediction accuracy, we compare a set of most promising topics as returned by our predictive model to real-world most promising topics.
2. **Growth** Promising topic prediction in the sense of growth is achieved through the use of a regression model to predict either ΔDFT or ΔDAF feature of topics. Further steps are the same as described above.

7 Experiments

7.1 Experimental setup

1. **Scope** For our experiments, we have investigated three periods, starting with the value of Y_1 as 2007, 2008 and 2009. In each of these cases, n was

equal to 5. What it does mean for the model is, that we train the regression model based on years 2007-2011 predicting year 2012, then test the model by using years 2008-2012 to predict year 2013. We perform analogous task for the remaining two values of Y_1 .

2. **Regression model and features** For the purposes of this experiment, we have chosen a linear regression model with feature selection. This choice was motivated by the ease of interpretation provided. There was a total of 28 input features prior to the feature selection process: DFT , DAF , $TF-IDF$ and $Popularity$ present five times each, once for every year, while ΔDFT and ΔDAF were present four times each calculated for every consecutive pair of years in the input data set.
3. **Verification** To verify the accuracy of our predictions, we rank the set of articles published in year Y_{n+2} based on the predicted value of a desired feature (for example, DFT) and the real value of the same feature. We then select top x positions from both rankings and acquire the common part of these two sets. Note that such operation disregards the precise order of the ranking. Our rationale is that it is more valuable to accurately predict that a specific topic will be among the top 20 in the following year, than whether it will be 5th or 10th. The final accuracy score is the proportion of the number of topics predicted to be in the top x compared to real top x topics.

7.2 Results

1. **Bag of words** The bag of words model achieved a reasonable accuracy when predicting promising articles by the value of their DFT feature, as shown in Fig. 3a. Note that the accuracy is much lower for year 2015 than the two preceding years. This might be related to the fact that DBLP database did not include arXiv articles prior to 2015. A sudden change in the volume of documents and, presumably, the distribution of words in their titles would be expected to lower the predictive ability of the model. Interestingly, prediction focused on the size of a community (DAF , Fig. 3b), remained unaffected. The situation is different for the case of prediction based on the community growth (Fig. 3d) as it suffers from lesser consistency prior to the year of 2015. Growth measure in regards to the number of documents published on a given topic is predicted with lower variation of accuracy (Fig. 3c). In both cases the predictive model fares much worse than when operating on raw numbers.
2. **LDA** Predicting promising research topics in the sense of a significance yields very accurate results, as seen on Fig. 4a and Fig. 4b. This aligns with the results achieved through the use of a bag of words model. In the case of both DFT and DAF feature, the linear regression model is capable of good predictions that remain consistent within the scope of a year. We do not observe either of the features to be predicted with visibly lower accuracy, unlike in the previous case. This time, however, the year of 2014 scores the lowest across the lengths of rankings and features as well. This accuracy drop is not as dramatic as in the case of bag of words model, but consistent enough

to be visible. The growth metrics, are predicted with a significantly lower accuracy, evidenced on Fig. 4c and Fig. 4d. Attempts of predicting any of the two features considered as indicative of growth fail to produce consistent results; for each of the three publishing years taken into consideration, it is either the prediction of a ΔDFT or ΔDAF feature that achieves varied accuracy between the length of a ranking considered. In some cases the most discriminating subset, top ten research topics, scores 0% accuracy. This is considerably worse than the results of the same task carried out with the use of a bag of words model.

Fig. 3: Prediction accuracy using the bag of words model

7.3 Discussion

The stark difference between results of our predictions for raw numbers and the growth of such numbers can be explained intuitively. As mentioned before, once a topic achieves a wide following, it is likely to remain a sizable research area for a long time. This introduces little dynamic to the top rankings, as they can be expected to bear similarity to each other between the years. Changes in numbers,

on the other hand, are relative and capture more information. Thus, it would be expected that a task of such prediction would be harder.

Fig. 4: Prediction accuracy using the LDA model

The metrics available within the database are crucial. When stripped of purely bibliographical information, the DBLP data set contains little quantifiable information; features proposed and included in our research were computed based on the records in the data set. As shown by the predictive results presented, operating on basic features might yield results when attempting to predict other simple features, but is insufficient to capture true nuance. In such a case, the benefit of high interpretability is of comparably little value.

It is impossible to discuss the results of our research without retrieving the most influential bi-grams and research topics from our topic identification models. Accuracy scores can be high, but a manual review of the supposed most promising research areas is necessary. For this reason, we present two most promising topics retrieved by LDA model, with regards to Δ DFT on Figure 5 and a list of five most promising bi-grams, by the same metric, in Table 2.

Topics retrieved by the LDA model appear anomalous; while there seems to be an internal connection between the words, these are either not hinting towards

a realistic area of research or are mixed with one or more similarly related groups. A notable problem is the abundance of words common in computer science, such as "system" or "application". These would belong to a list of stop-words in an ordinary case. However, through the course of several experiments we have determined that no matter the extent of a stop-list, there would be a stable number of about 20 words assigned to anywhere between 5 and several hundred topics. This leads us to the conclusion that it is extremely hard to retrieve topics from titles of scientific articles alone. This seems in line with what we can observe in scientific literature, as the titles shifted from being purely informative towards attracting the potential reader with a unique description of the problem.

While bi-grams are more suited for identifying research areas from titles, they also prove to have limitations. Topics like "deep learning" can be artificially over-represented, as they commonly attract additional attention to the work. Unlike more elaborate mechanisms, using a BoW approach doesn't allow for identifying a topic when nuanced or metaphorical wording is used during the title-creation, leading to inserting one or more words into an established phrase.

8 Conclusions

Rank	Topic
1	deep learning
2	data analytics
3	deep neural
4	massive system
5	full duplex

Table 2: Top 5 most promising bi-grams with respect to Δ DFT

We have presented our approach for simple and interpretable identification of promising topics in computer science. The approach itself shows promise, but seems to require more nuanced and in-depth features to yield high-value results. We have highlighted that, for the use of LDA model, titles of scientific publications cannot be treated as any other short text. We believe this is caused by scientific community's growing awareness of the benefits that higher marketability brings in.

In the future work, we will aim at acquiring a data set containing advanced information to measure its impact on our predictive abilities. Furthermore, we are interested in measuring in higher detail the impact a community working on a scientific topic can have on the topic itself.

Fig. 5: Word clouds of top two topics retrieved by the LDA model

Acknowledgement

This work was partially supported by the National Science Centre, Poland, project no. 2016/21/B/ST6/01463 and by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant no. 691152 (RENOIR); by the Polish Ministry of Science and Higher Education fund for supporting internationally co-financed projects in 2016-2019, no. 3628/H2020/2016/2.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Advances in neural information processing systems* 1, 601–608 (2002)
2. Glänzel, W., Thijs, B.: Using ‘core documents’ for detecting and labelling new emerging topics. *Scientometrics* 91(2), 399–416 (2012)
3. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, L.: Detecting topic evolution in scientific literature: How can citations help? In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. pp. 957–966. CIKM '09 (2009)
4. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: *Proceedings of the First Workshop on Social Media Analytics*. pp. 80–88. SOMA '10 (2010)
5. Hurtado, J.L., Agarwal, A., Zhu, X.: Topic discovery and future trend forecasting for texts. *Journal of Big Data* 3(1), 7 (2016)
6. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. pp. 63–70. ETMTNLP '02 (2002)
7. Lu, Y., Zhang, P., Liu, J., Li, J., Deng, S.: Health-related hot topic detection in online communities using text clustering. *PLOS ONE* 8 (02 2013)
8. Mund, C., Neuhäusler, P.: Towards an early-stage identification of emerging topics in science—the usability of bibliometric characteristics. *Journal of Informetrics* 9(4), 1018 – 1033 (2015)
9. Prabhakaran, V., Hamilton, W.L., McFarland, D.A., Jurafsky, D.: Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In: *ACL* (2016)
10. Wang, Y., Joo, S., Lu, K.: Exploring topics in the field of data science by analyzing wikipedia documents: A preliminary result. *Proceedings of the American Society for Information Science and Technology* 51(1)
11. Zhang, B., Guan, X., Khan, M.J., Zhou, Y.: A time-varying propagation model of hot topic on {BBS} sites and blog networks. *Information Sciences* 187, 15 – 32 (2012)