DOI: http://dx.doi.org/10.14236/ewic/FDIA2017.11

Trend Mining on Bibliographic Data

Christin Katharina Kreutz
Trier University
54286 Trier, DE
kreutzch@uni-trier.de

Most times, current and emerging trends in scientific communities are only discernible for active domain experts. Trend mining on bibliographic data may give insight for politicians, entrepreneurs and new scientists. This work outlines a research agenda for approaching this problem.

Trend Mining, NLP, Machine Learning

1. INTRODUCTION

The evolution of topics in a scientific community and terms associated with them is something important to intensively follow for researchers because it may affect their work. For politicians who want to fund projects, the industry which wants to adopt current findings and new scientists it is difficult to determine what is trending right now. A topic is a trend if it is not popular for a certain time or not existing but its significance soars. This change could be signalled by *important* researchers starting to work on this theme or its appearance in seminal conferences and journals. Automatic detection of these trends (*Trend Mining*) with special focus on and usage of the underlying network of authors, publications, subdivisions in the fields and their connections is a big challenge.

The dblp computer science bibliography contains bibliographic information related to publications, authors, conferences and journals from the field of computer science and adjacent areas (Ley (2009)). As of August 2017, it stores meta-information concerning over 3.8 million publications and 1.9 million authors. For about half the publications, additional data like abstracts is available as a supplement to dblp. For some publications, keywords and citations are in stock. The textual information and citation data is currently being extended by open access collections. This research aims at trend mining based on this bibliographic data set.

There are different types of trend mining which are relevant for the problem at hand: *i) blog mining*, which works on blogs and takes fickle and unstandardised associations between authors (lists of authors' favourite bloggers, lists of referrers to entries, mentions or linkings in blog entries) into account (Glance et al. (2004)), *ii) social network mining*, which analyses short-term connections (movement of cattle) in social networks but is geographical dependent (Nohuddin et al. (2016)) and *iii) spatiotemporal mining*, which observes the connection and evolution of topics through added labels by users on a geographical level (He et al. (2016)).

Other publications exist, where trends are substantiated with the prediction of keyword distributions on textual information but no underlying structure is taken into account (Asooja et al. (2016)). Further influential approaches which will be relevant for this work are the observation of citations as well as textual information of publications in the definition of emerging topics (Glänzel et al. (2012)) and the detection of new themes as they are coming up with the help of a topic network deployed from the relations between their descriptive keywords and publications, authors, venues and organisations (Salatino et al. (2016)).

However, none of the existing works utilizes the full spectrum of information available from a bibliographic data set to identify trends and the corresponding seminal persons, journals or conferences.

A goal for this research is to exploit the backbone of formal and therefore refined and reliable connections of authors through co-authorships, cited papers, conference participations and entries in the same journals to create a strong information base. It can be combined with textual information from publications for mining trends over big spans of time, if this data is available. In addition, word embeddings can be

used to improve the semantic interrelation of terms (Mikolov et al. (2013)). The resulting system should present trends and predict upcoming developments.

2. RESEARCH DIRECTION

This work intends to perform trend mining on bibliographic data and to find future directions in which research may develop. The building blocks of this prediction contain the identification of important researchers and seminal conferences or journals. These researchers could be described as influential persons which dominate an area and are central in a topic. They are cited numerous times and advance science. Pathbreaking conferences and journals can be found by observing the history of trends and identifying those where these themes appeared in right before they became popular.

A model for the evolution of topics is essential for analysing the breeding grounds for trends and can be achieved in numerous steps: *1) understanding* temporal dynamics of research themes by precisely observing past development of trends, *2) correlating* previous trends via machine learning with bibliographic features, *3) applying* the gained knowledge onto a current time frame to find prevailing trends and predict upcoming ones.

The influence of different features from bibliographic metadata is a further aspect of this work. While semantically-annotated versions of full texts of publications should distinctly improve the mining process, these texts are unavailable at most times. This work then aims to compensate for such information deficits by using component analysis on citation- or co-author- graphs augmented by available information on content from titles, abstracts or full texts.

As a part of this, the effect of using methods for cleaning and preparing the data including stemming, a weighting of data fragments concerning their influence and semantic annotation of textual content will be examined.

There are two sides the system operates on: past and prediction. For the past, the resulting system will retrieve central persons, important publications as well as seminal journals and conferences for a given topic. The prediction is performed by the computation of future trends with specification of their respective indicators such as influential papers and pathbreaking conferences and journals.

While trend mining is the primary focus of this research, many other applications of the underlying data and building blocks can be imagined.

With this information at hand, a recommender system could be constructed to propose fitting publications to discuss in the ”Related Works”-section of a paper. Additionally, such a system could suggest noteworthy but unmentioned references to important works by analysing the full text and existing citations.

A further application of the system could be an extension for recommendations of suitable reviewers from the same field for an unreleased publication submitted to a conference or journal.

Existing recommendation systems tend to focus on word-topic, topic-citation distributions and concentrate on suggesting fitting but not necessary influential publications (Huang et al. (2014)) or are only applicable for already established conferences as they heavily rely on past program committees of the same conference in their suggestion of experts for current program committees (Tran et al. (2017)). This implementation has the potential to be more precise than existing work because it capitalises from all the surrounding information which is normally not used in these systems and it does not depend on strong assumptions.

Although further data sources such as Twitter are available and might be beneficial in the detection of changes in themes, this work focuses solely on bibliographic information.

3. EVALUATION PLAN

The evaluation of the approach will be performed with a sliding-window-frame on the temporal axis of data from dblp and results will be cross-validated.

The first move of the evaluation process would be finding and interviewing experts from different domains in computer science, to compose a list of trends with corresponding years. The resulting enumeration is then used to train the system with data up to a point in time. For the next step, the system predicts future trends that may emerge after this interval and how research would evolve in a certain time-frame, i.e. in the following five years. A last action is the comparison of findings from the automatic approach to the list the experts deployed for this window.

Aside from this, the forecasted trends of the system which were not contained in the human-generated list should be returned to the experts. They then have to determine the degree of falseness of the results as they could be completely off themes, minor topics or trends whose evolution has been cut off by unpredictable events. An attempt will be performed to find automatic methods for this assessment.

The feature contribution from bibliographic metadata as well as the different techniques in preprocessing the data will also be analysed with regard to efficiency, required space and correctness of the constructed approach.

ACKNOWLEDGEMENTS

Special thanks goes to my supervisor Ralf Schenkel for his invaluable support. I thank the PROMOS-program of the German Academic Exchange Service (DAAD) for the awarded scholarship for participating in ESSIR 2017 and the FDIA Symposium. Finally I would like to thank the reviewers for their comments and suggestions to improve the quality of this work.

REFERENCES

Ley, Michael (2009) DBLP - Some Lessons Learned. *PVLDB,* 2(2). 1493-1500.

Glance, Natalie S. and Matthew Hurst and Takashi Tomokiyo (2004) BlogPulse: Automated trend discovery for weblogs. In: *WWW '04.*

Nohuddin, Puteri N. E. and Frans Coenen and Rob Christley (2016) The application of social network mining to cattle movement analysis: introducing the predictive trend mining framework. *Social Network Analysis and Mining.* 6(1). 45:1-45:17.

He, Jiangen and Chaomei Chen (2016) Spatiotemporal Analytics of Topic Trajectory. In: *VINCI '16. Dallas, 24-26 September 2016.* New York: ACM. 112-116.

Asooja, Kartik and Georgeta Bordea and Gabriela Vulcu and Paul Buitelaar (2016) Forecasting Emerging Trends from Scientific Literature. In: *LREC 2016. Portorož, 23-28 May 2016.* ERLA.

Glänzel, Wolfgang and Bart Thijs (2012) Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics,* 91(2). 399-416.

Salatino, Angelo Antonio and Enrico Motta (2016) Detection of Embryonic Research Topics by Analysing Semantic Topic Networks. In: *SAVE-SD 2016. Montreal, 11 April 2016.* Cham: Springer International Publishing. 131-146.

Mikolov, Thomas and Kai Chen and Greg Corrado and Jeffrey Dean (2013) *Efficient Estimation of Word Representations in Vector Space.* Available from https://arxiv.org/pdf/1301.3781.pdf (20 July 2017).

Huang, Wenyi and Zhaohui Wu and Prasenjit Mitra and C. Lee Giles (2014) RefSeer: A citation recommendation system. In: *JCDL '14. London, 8-12 September 2014.* Piscataway: IEEE Press. 371-374.

Tran, Hong Diep and Guillaume Cabanac and Gilles Hubert (2017) Expert suggestion for conference program committees. In: *RCIS 2017. Brighton, 10-12 May 2017.* IEEE. 221-232.