A Bimodal Network Approach to Model Topic Dynamics

Di Caro, L.; Guerzoni, M.; Nuccio, M.; Siragusa, G.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

A Bimodal Network Approach to Model Topic Dynamics

Luigi Di Caro*, Marco Guerzoni**, Massimiliano Nuccio**, Giovanni Siragusa*

* *dicaro@di.unito.it; siragusa@di.unito.it*
Department of Computer Science, University of Turin, Via Pessinetto 12, Turin, 10149 (Italy)
Despina Big Data Lab (Lab or Department), Lungo Dora Siena, Turin (Italy)

** *marco.guerzoni@unito.it; massimiliano.nuccio@unito.it*
Department of Economics and Statistics Cognetti de Martiis, University of Turin, Turin, 10149 (Italy)
Despina Big Data Lab (Lab or Department), Lungo Dora Siena, Turin (Italy)

Introduction

A crucial issue in the philosophy of science consists in the understanding of the evolution of scientific paradigms within a discipline. Following Kuhn [1970, p.10], a scientific paradigm can be thought as the set of assumptions, legitimate theories, methods, and experiments both adequately new to attract a group of scholars, to build a contribution to a field and to open enough the exploration of different directions of research.

In the traditional view, as developed for hard and mature sciences, the evolution of scientific paradigm consists in "the successive transition from one paradigm to another via revolution [Kuhn, 1970, p.12]. However, a scientific field is usually composed by several research paradigms either competing or addressing different issues, and a revolution in one of those necessarily involves effects and readjustments in the entire discipline. Moreover, each new paradigm carries the legacy of the existing knowledge of past paradigms, which is often recombined into the new one. This is especially true for social sciences, in which the identification of clear scientific paradigms in the sense of Kuhn is often blurred and it is probably more correct referring to "research traditions" [Laudan, 1978].

However, whether you call paradigms or traditions, the existence of patterns of thoughts which are legitimate contributions to a theory is undeniable. Thus, we can postulate that the evolution of knowledge in a scientific field is generated among a community of researchers which share a semantic area to define specific research issues, describe methodologies, and lay down results. Thus, the heterogeneity of the research tradition of a scientific field can be described with semantic analysis.

The idea that some measure of words co-occurrence reveals an underlying epistemic pattern and, therefore, it can capture the essence of evolution in science is not a new one. Despite the difficulty in programming, the first attempts date back to the work of Callon et al. [1983] and refined when the first open code have been made available a decade later [Vlieger and Leydesdorff, 2011, Leydesdorff and Welbers, 2011].

The challenge of classifying science on the basis of its semantic content has found a renewal with the diffusion of machine learning techniques and, in particular, in the subfield of unsupervised learning [Leydesdorff and Nerghes, 2015]. Topic modeling includes a family of algorithms [Blei et al., 2003], which are particularly performant in extracting information from large corpora of textual data by reducing dimensionality. This feature has been clearly recognised in mapping science [Suominen and Toivanen, 2015] or news [DiMaggio et al.,

2013]. Alghamdi and Alfalqi [2015] review four major methods of topic modeling, including Latent Semantic Analysis (LSA), Probabilistic LSA, Latent Dirichelet Allocation (LDA) and Correlated Topic Model (CTM). The LDA proposed in [Blei et al., 2003] is one of the most diffused approaches. LDA retrieves latent patterns in texts on the basis of a probabilistic Bayesian model, where each document is a mixture of latent topics described by a multinomial distribution of words. One of the major limitations of LDA lies on its inability to model and represent relationships among topics over time [Alghamdi and Alfalqi, 2015].
In this paper, we address a major recurring issue in topic modeling, that is the topic dynamics, or, in other words, we test a method to track the transformation of topics over time. As stated by Blei and Lafferty [2006], LDA is a powerful approach to reduce dimensionality, but it assumes that documents in a corpus are exchangeable. On the contrary, articles and themes are sequentially organized and evolve over time. Therefore, it is not only relevant to develop a statistical model to determine the evolving topics from a corpus of a sequential collection of documents, but also to measure and describe the transformation of topics and their appearance and disappearance.
In the literature of information retrieval, the dynamics of topics has been faced with two approaches [He et al., 2009]: a discriminative one monitors a change in the distribution of words or in the mixture over documents, while a generative approach searches for general topics over the whole corpus and, then, it assigns the documents which belong to each topic [Bolelli et al., 2009, He et al., 2009].
Specifically Blei and Lafferty [2006] introduced Dynamic Topic Modeling (DTM), a class of generative models in which the per document topic distribution and per topic word distributions are generated from the same distributions in a previous time frame. This approach has been very influential since it imposes a connection between the sets of topics at different periods and allows to track the evolution of a single topic over time. DTM performs very well in capturing the evolution of a single topic. However, the evolution of knowledge is much more complicated that the change of relative importance of words within a topic, since it may involve also the creation of new topics, their mutual re-combinations and, eventually their possible demise. The major contribution of the paper is the conceptualization and formalization of the evolution of knowledge, conceived as different streams of semantic content which continuously appears and disappears, merges and splits. Thereby we propose an original method based on inter-temporal bimodal networks of topics compute the key elements in the evolution of knowledge.
Moreover, the ultimate goal of the paper is not to track in detail what happens within a single topic, but rather to develop indexes which can measure at the aggregate level some properties of the observed knowledge dynamics, such as an overall degree of novelty or the level of turbulence at specific time windows.

A Conceptualization of Topic Evolution

In this paper, we focus on the dynamic evolution of topics over time. With DTM, each topic Kt is linked to Kt+1 creating a topics chain which spans the years covered by the documents. Specifically, Blei and Lafferty [2006] maps each topic at time t-1 into a topic in t by chaining the per document topic distribution and the per topic word distribution in a sate space model with a Gaussian noise. This approach is highly performing to track incremental changes of the same topic but it does not focus on revealing neither birth nor death nor possible combinations of topics and it imposes a constant number of topics within the model. On the

contrary, we are interested to discover the structural change of topics in a corpus and to understand the underlying topic dynamics which explain it. Thereby, we do not focus on the evolution of the single topic. The inter-temporal link across topics is not a constraint in the estimation of the model as in the DTM, but it is introduced ex-post in the empirical analysis by looking at the similarities (co-occurrence of words) amongst topics generated by independent LDAs. More in detail, while DTM models sequences of compositional random variables by chaining Gaussian distributions (thus directly embodying topics dynamics in the model), our approach operates on single and static LDAs in order to track and measure such dynamics out of the model.

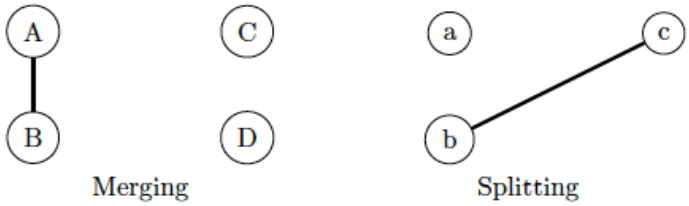
The evolution of a topic structure of a corpus accumulating knowledge overtime takes place because of two main reasons. On the one hand, any epistemic community (say for instance journalists or scientists) can shift their intellectual interest to new issues and problems, which will result in different choices, frequencies and co-occurrence of words. On the other hand, language is subject to a constant evolution, in which new words, named entities, acronyms, etc. appear while other ones disappear due to an increasingly lesser use of them by the same community. We rule out this second scenario, by assuming that in the short time frame the language is fairly stable.

Under this assumption, when comparing the topics generated by a topic modeling exercise in two different, although adjacent, time windows, we should be able to capture the evolution of the scientific debate and highlight the birth, death and recombination of topics. On the one extreme, we can find a situation in which knowledge does not evolve and thus topics are stable. On the other, we figure out the maximum of turbulence in which new topics emerge without any semantic relation with the incumbent ones. In the latter case, we may assume the death of past topics and the birth of new ones. In between the two ideal cases, we can also draw a continuum in which we can observe both deaths and births of topics. Finally, in a most interesting scenario, rather than observing stability or turbulence, knowledge may evolve recombining existing topics in both old and new ones.

Let us consider M topics emerged as the result of a topic modeling exercise from a corpus of articles at time t and N topics at time t+1. We tackle the critical problem of tracking the transformation of the set of topics M at t into the set of topics N at t+1. Specifically, we are interested in measuring the magnitude of the various phenomena such as birth, death, merging, and splitting. Consider a similarity index based on word co-occurrence between each couple of topics and consider the similarity matrix S (M x N). Births and deaths can be easily calculated from the matrix S. A row sum equal to zero highlights a death, while a column sum equals to zero indicates a birth. A death means that the semantic legacy completely disappears while a birth means that a topic carries no semantic similarity with other topics in the past. Once again it is important to notice that these cases are extreme scenarios while in the reality we observe a continuum between births and deaths. We might thus calculate an index Novelty (NI) for each topic i at time t+1 where for NI = MAX we have a birth, that is a topic with no similarity to any other previous one. For higher value we have a higher novelty of the topic. We can also measure an average change in NI on the overall structure of a scientific field by looking at distributions of these indexes over the topics. We take the average of all the cell values in matrix S. If the similarity index is bounded between 0 and 1, such it is the very common case of the cosine similarity index, thus NI ranges from 0 to 1. For very small value of novelty, new topics show different word distribution from old ones. As mentioned, transformation of topics can take the form of

merging and splitting. We say that a merging occurs if a topic at time t+1 shows a high similarity with two topics at time t, meaning that the semantic universe of A and B at t is combined in the topic a. Similarly, we can say that a split occurs if the semantic legacy of one topic at t is to be found in multiple topics at t+1 as in the case for topic C. To analyse the intensity of a merging we can project the bipartite network into its two 1-mode-network. This is achieved by a matrix multiplication S x S(transposed) for the merging and S(transposed) x S for the splitting which result in two matrices P-merging and P-splitting of dimension respectively M x M and N x N.

Figure 1: Mode network.



In this way, we can compute a MergingIndex (MI) which takes value 0 when no merging occurs and it ranges up to an upper limit which can not exceed 1.

$$P^{merging}_{normalized} = P^{merging} \cdot \frac{1}{\sum_{i \leq j} P(i,j)} \qquad MI = 1 - trace(P^{merging}_{normalized})$$

Symmetrically, we calculate a SplittingIndex (SI).

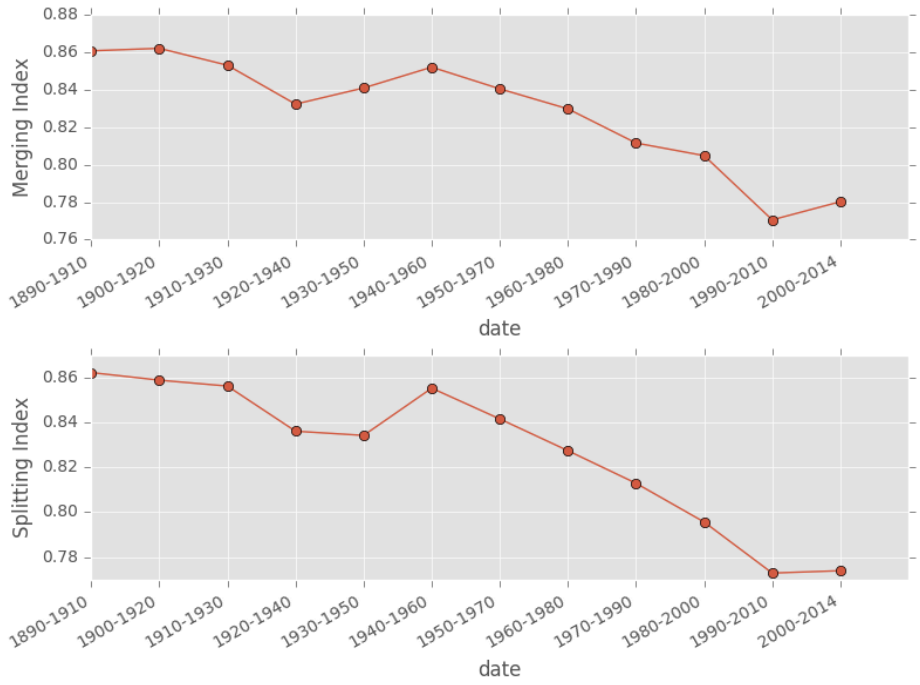
Experimentation

The used dataset is a collection of documents which appear in the JSTOR database (www.jstor.org)and were published from 1845 to 2013 in more than 190 journals concerning with economic sciences (also defined as economics). They are more than 460,000 documents, classified as research articles (about 250,000), book reviews (135,000), miscellaneous (73,000), news (4,000) and editorials (500). For each document, in addition to bibliographic information (title, publication date, authors, journal title, etc.), the dataset provides full content in form of a bag of words, i.e. the set of words used in the documents associated with their frequencies.

The LDA has been applied to research papers published between 1890 and 2013: decades before 1890 were dropped because of the extremely low number of documents. Thereby, the resulting dataset of articles consists of 755,838,336 words and 3,169,515 unique words. We experimented varying the hyper-parameters of the method, namely the number of topics and the dimension of time windows, in order to evaluate the robustness and sensitivity of our approach in the 123 years considered. We selected 25, 50 and 100 topics and time windows of 5, 10 and 20 years, keeping fixed one parameter and varying the other one. In details, we first analyzed the values of SI and MI fixing the window dimension to 10 years and varying the number of topics. These simple tests demonstrated that the main trends of the indexes do not change substantially by varying the hyper-parameters, meaning that our method is robust to the number of topics and the size of the time windows. Figure 2 shows the values of MI and

SI respectively for each time window. In the used corpus, both indexes show a general trend of decreasing values over time, which becomes particularly severe starting from 1960s. Merging and splitting increase only between the 1940s and the 1950s while dropping dramatically in the second half of the XX century. The transformation of topics seems to find new urge only around the end of the century, when merging is increasing again and splitting is stable.

Figure 2: Obtained MI and SI.



Conclusions

In this paper we proposed a method to measure the evolution of knowledge in a scientific field extracting topics in a corpus of documents. Topic modeling techniques are becoming increasingly refined in treating large and complex corpora of documents, but they may lack of a theoretical reflection of the underlying empirical phenomenon. Taking a dynamic perspective we recognise five paradigmatic cases of knowledge evolution. We then surmise that modeling the proximity between topics of different time windows as a proximity network might be a useful tool to measure their knowledge dynamics. Indeed, this network approach allows us to develop 3 indexes, which grasp i) the stability of topics over time measuring their rate of death and birth, and ii) the degree of recombination of topics. For very simple cases, we are also able to analytically derive those conditions, which link the proximity network and the value of each index. Testing the algorithm over a set of simulated documents, we showed its robustness for each the indexed developed. We believe, this is a first step towards the development of a closer connection between algorithms for dynamic topic modeling and the empirical phenomenon they are supposed to describe.

References

R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. International Journal of Advanced Computer Science and Applications , 6(1):147–153, 2015.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning , ICML '06, pages 113–120, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2..

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research , 3:993–1022., January 2003.

L. Bolelli, S. Ertekin, D. Zhou, and C. L. Giles. Finding topic trends in digital libraries. In Proceedings of the 9th ACMIEEE-CS joint conference on Digital libraries , pages 69–72. ACM, 2009.

M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin. From translations to problematic networks: An introduction to co-word analysis. Social science information , 22(2):191–235, 1983.

P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. Poetics, 41(6):570–606, 2013.

Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. Detecting topic evolution in scientific literature: how can citations help? Proceedings of the 18th ACM conference on Information and knowledge management , pages 957–966, November 2009.

T. S. Kuhn. The structure of scientific revolutions, International Encyclopedia of Unified Science, vol. 2, no. 2 . Chicago: The University of Chicago Press, 1970.

L. Laudan. Progress and its problems: Towards a theory of scientific growth . Univ of California Press, 1978.

L. Leydesdorff and A. Nerghes. Co-word maps and topic modeling: A comparison from a user's perspective. arXiv preprint arXiv:1511.03020 , 2015.

L. Leydesdorff and K. Welbers. The semantic mapping of words and co-words in contexts. Journal of Informetrics , 5(3):469–475, 2011.

A. Suominen and H. Toivanen. Map of science with topic modeling: Comparison of unsupervised learning and human?assigned subject classification. Journal of the Association for Information Science and Technology , October 2015.

E. Vlieger and L. Leydesdorff. Content analysis and the measurement of meaning: The visualization of frames in collections of messages. Public Journal of Semiotics, 3(1):28–50, 2011.