Advances Towards Early Detection of Research Topics

Angelo Antonio Salatino
Knowledge Media Institute
Open University
Milton Keynes, UK
angelo.salatino@open.ac.uk

Supervisors names: Prof. Enrico Motta, Dr. Francesco Osborne
Status: Full-Time
Probation viva: After
Starting date: 01/12/2014

Keywords— Scholarly Data, Research Trend Detection, Topic Emergence Detection, Topic Discovery, Semantic Web, Ontology

Acknowledging new trends in the research environment is important for many stakeholders, such as researchers, institutional funding bodies, academic publishers, and companies. In particular, being able to identify them as soon as possible can bring an important strategical advantage.

A trend is usually defined as the general direction in which something is evolving. It is often used to describe the popularity of items, such as brands, words, and technologies. In order to detect trends, the relevant items should usually be already recognized and often somewhat popular. For this reason, current methods for detecting trends of research topics usually focus on identifying terms associated with a substantial number of documents, which usually took some years to be produced. Conversely, I theorise that it is possible to perform very early detection of research trends by identify embryonic topics, which have not yet been explicitly labelled or identified by a research community, and that is possible to do so by analysing the dynamics of existent topics. My work is grounded in Kuhn's theory [1] of the scientific revolution according to which a paradigm shift, also called scientific revolution occurs when a paradigm cannot cope with anomalies, leading to a crisis that will persist until a new outcome redirects research through a new paradigm. In this abstract, I will discuss the state of the art, present an initial study which supports my hypothesis and outline the future directions of my work.

The state of the art presents several works regarding the detection of research trends which can be characterised either by the way they define a topic or by the approach they use to discover trends [2]. A common approach to identifying research topic is the probabilistic topic model, in which a topic is characterised as a multinomial distribution over words. The Latent Dirichlet Allocation (LDA) [3] is the most popular technique to extract topics from a corpus using this definition. Since its introduction, LDA has been extended and adapted in many other applications as it can be seen in He et al. [4], in which LDA and citation networks were combined to address the problem of topic evolution. In some other approaches, keywords are used as proxies for research topics [5], but Osborne et al. [6] pointed out several drawbacks in using keywords, since they tend to be noisy and carry an unexploited implicit relation between themselves. Further approaches [7, 8] used taxonomy of topics, which can provide a better characterisation of topics and contain semantic relationships between research areas.

The approaches for detecting research trends usually rely on statistical techniques to analyse the impact of labels or distributions of words associated to topics [4, 5, 8]. However, all these approaches focus on already existing topics which are usually already associated with a substantial number of publications.

Considering the current gap, it is legitimate to ask *"How is it possible to detect the early emergence of new research topics?".* Thanks to the availability of very large repositories of scholarly data, nowadays it is possible to address this question.

To confirm my thesis, I investigated whether the emergence of a new topic is anticipated by specific dynamics among existing ones. To do so, I introduced a novel method for assessing the increase in the pace of collaborations in topic networks and tested it on more than 2000 co-occurring topics and 3 million research publications from the Rexplore system: http://technologies.kmi.open.ac.uk/rexplore. In particular, I randomly selected 50 topics that emerged in the decade 2000-10 for my treatment group (debutant group) and 50 well established topics as a control group (non-debutant group). All these topics were selected within the domain of Computer Science and defined according to the taxonomy produced by Klink [6]. The experiment itself consisted in two phases. In the first phase, I selected and extracted portions of the collaboration network that was related to the topics in the two groups in the few years prior the year of their debut (the topics in the control group were associated to random years). Afterwards, I analysed the overall pace of collaboration for each network associated to these test topics.

Fig. 1 shows the steps of the experiment. The selection phase is based upon the assumption that an emerging topic will tend to collaborate with its procreators. Therefore, these topics could be analysed in the previous years to confirm my theory. Hence, for each topic in the two groups, I selected their *n* (20, 40, 60) most co-occurring ones and then extracted the portion of their collaboration network containing these topics in the five years prior to the year of analysis. A collaboration network is a fully weighted graph in which nodes are represented by topics and their weight represent the number of papers in which they appear, while links between nodes and their weights represent the amount of papers they co-occur together. At the end of this stage, each tested topic was associated with five collaboration networks, representing the behaviour of its predecessors in the five previous year. I then measured the increase in the pace of collaborations in this network with a number of metrics, which involved the analysis of 3-cliques, that are apt to model small collaboration between nodes.

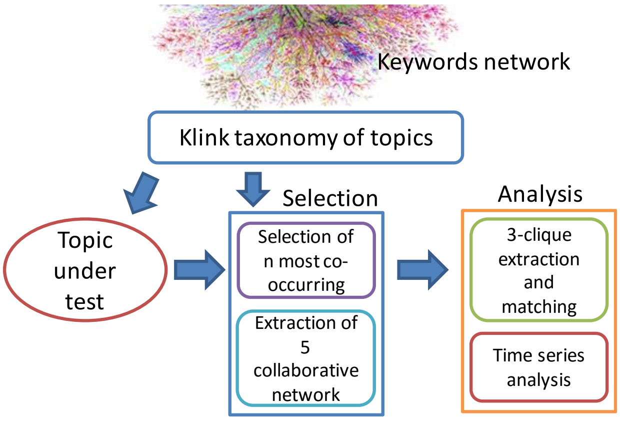
Keywords network
Klink taxonomy of topics
Topic under test
Selection
Selection of n most co-occurring
Extraction of 5 collaborative network
Analysis
3-clique extraction and matching
Time series analysis

Fig. 1. Example of framework applied to one of the topics under test

As reported in my previous work [9], the finding shows that the portion of network that is related to a debutant topic exhibits a higher pace of collaboration than the portion of network related to non-debutant topics. In particular, by using an approach based on the linear regression of the collaboration pace associated to each year it is possible to effectively discriminate the two groups of topics. In addition, by analysing the collaboration networks containing 20, 40 and 60 most co-occurring topics, I found out that increasing the size of collaboration networks, the approach provides better results. A reason for this can be that increasing the number of topics in the collaboration network increases the chances to select its procreators. I performed the Student’s t-test over the distributions associated with the two groups, obtaining p-values less than 0.0001, which allows me to reject the null hypothesis H_0: *“The differences in the pace of collaboration between the debutant topics and topics in the control group result purely from chance”*.

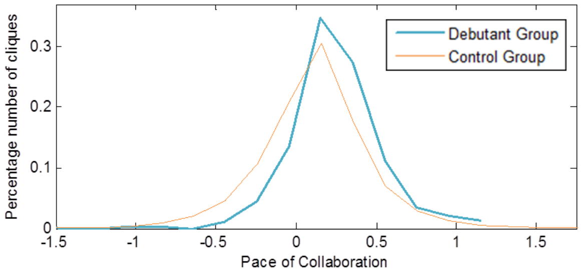

Fig. 2. Distribution of pace of collaboration for both groups

Fig. 2 shows the two distributions of pace of collaboration obtained for the debutant group (blue) and non-debutant groups (orange), in which is possible to see that the distribution of the debutant group is shifted towards positive values while the distribution of the control group is almost centred in zero.

In conclusion, the results of the preliminary analysis confirm my initial hypothesis, i.e., that it should be possible to anticipate topics emergence by analysing the dynamics between existent ones. I plan to further develop my approach in two main directions. First, I am currently working on a method for the automatic detection of embryonic topics that analyses the topic network and identifies sub-graphs where topics exhibit the discussed dynamics. A second direction of work focuses on improving the current approach by integrating a number of additional dynamics involving other research entities, such as authors and venues. The aim is to produce a robust approach that could be used by researchers and companies alike for gaining a better understanding of where research is heading.

References

[1] T. S. Kuhn, *The structure of scientific revolutions*: University of Chicago press, 2012.

[2] A. Salatino, "Early Detection and Forecasting of Research Trends," 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993-1022, 2003.

[4] 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*, pp. 957-966, 2009 2009.

[5] A. Duvvuru, S. Radhakrishnan, D. More, S. Kamarthi, and S. Sultornsanee, "Analyzing Structural & Temporal Characteristics of Keyword System in Academic Research Articles," *Procedia Computer Science*, vol. 20, pp. 439-445, 2013.

[6] F. Osborne and E. Motta, "Klink-2: integrating multiple web sources to generate semantic topic networks," in *The Semantic Web–ISWC 2015*, ed: Springer, 2015, pp. 408-424.

[7] F. Osborne, G. Scavo, and E. Motta, "A hybrid semantic approach to building dynamic maps of research communities," in *Knowledge Engineering and Knowledge Management*, ed: Springer, 2014, pp. 356-372.

[8] S. L. Decker, B. Aleman-Meza, D. Cameron, and I. B. Arpinar, "Detection of bursty and emerging trends towards identification of researchers at the early stage of trends," University of Georgia, 2007.

[9] A. A. Salatino and E. Motta, "Detection of Embryonic Research Topics by Analysing Semantic Topic Networks," 2016.