

## Internship proposal (ERIC Lab – University of Lyon, France)

**Master Thesis:** Evolution of scientific domains: a longitudinal analysis

**Supervision:** J. Velcin, S. Loudcher

**Place:** ERIC Lab (University of Lyon)

**Duration:** 6 months

**Funding:** ERIC Lab

### Subject:

The Web makes the results and discoveries obtained through the scientific progress much more available nowadays than in earlier decades. In particular, scientific articles are used by worldwide teams not only for pursuing their own interests, but also for starting new collaborations, creating special journal issues, and organizing international conferences. These *events* seldom occur randomly: they are often the result of preliminary works and transversal collaborations. Researchers of the ERIC Lab are currently working on the longitudinal analysis of bibliographic data, with a particular interest in the diagnostics and prediction of such events.

In this context, the student's work will focus on the comparison of existing techniques for analyzing the evolution of the scientific literature. Different approaches can be tested and the student should make a full state of the art. This includes, on the one hand, the previous works dealing with bibliographic data [5, 7, 11]. On the other hand, she/he could see the topic models developed more especially for text mining [2, 4, 10]. These models are mainly focused on the extraction of the main tendencies but could be extended to deal with events [1, 8]. Several packages are freely available nowadays for testing these probabilistic models (for instance, see Mallet <http://mallet.cs.umass.edu>).

In order to validate the chosen approaches, the student will have access to various bibliographic databases, such as DBLP (<http://www.informatik.uni-trier.de/~ley/db>) or Pascal (<http://www.ovid.com/site/catalog/DataBase/214.jsp>). The comparison will be based on purely predictive quantitative measures, such as perplexity, but also on human judgment [3]. Recent work on topic extraction evaluation has been done in the lab [6, 9] ; it can be used as a start by the student.

The second stage of this internship consists in focusing on one interesting event: emergence of a new topic, publication of an original paper, creation of a new community, etc. To begin with, we propose to follow a purely descriptive approach in order to build an *explanation* of the occurrence of this particular event. The last stage will draw future lines of research for developing new machine learning models dealing with scientific events. These (future) models must be able to handle the different actors who populate a research domain: articles, authors, journals, conferences etc.

### References

- [1] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, 2002.
- [2] D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM New York, 2006.
- [3] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. volume 31, 2009.
- [4] A. Daud, J. Li, L. Zhou, and F. Muhammad. Conference Mining via Generalized Topic Modeling. *Machine Learning and Knowledge Discovery in Databases*, pages 244–259, 2009.
- [5] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley. Analysing social networks within bibliographical data. In *Database and Expert Systems Applications*, pages 234–243. Springer, 2006.

- [6] C. Musat, J. Velcin, M. A. Rizoïu, and S. Trausan-Matu. Improving topic evaluation using conceptual knowledge. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain, 2011.
- [7] M.C. Pham and R. Klamma. The Structure of the Computer Science Knowledge Network. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 17–24. IEEE, 2010.
- [8] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruíz-Shulcloper. Detecting events and topics by using temporal references. In *Proceedings of IBERAMIA 2002*, 2002.
- [9] M.A. Rizoïu and J. Velcin. *Topic Extraction for Ontology Learning*, pages 38–61. IGI Global, 2011.
- [10] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1), 2010.
- [11] O.R. Zaane, J. Chen, and R. Goebel. Mining Research Communities in Bibliographical Data? *Advances in Web Mining and Web Usage Analysis*, page 59, 2009.