

PhD proposal

Interactive analysis of phylomemetic structures

Zoltan Miklos (MCF), David Gross-Amblard (PR)
DRUID research group¹
Laboratoire IRISA, Université de Rennes 1

Funded by the ANR 2016 Epique project²

Keywords: evolution of science, analysis of very-large text corpora, mining dynamic networks, data quality, data exploration

Understanding the evolution of various scientific fields is important for our society. Obtaining a general picture of important evolutions of entire scientific fields is rather challenging in the light of the proliferation of scientific publishing and in the presence of overspecialized scientific journals. Recent papers [1,2] propose text analysis techniques to reconstruct important aspects of evolution, based on large corpora of scientific publications (such as Web of Science, PubMed).

The Epique project proposes to develop automated tools that can assist (social) scientists to study empirically particular aspects of the social dynamics of science. The existing methods for phylomemetic structure reconstruction rely on the following schema. 1) Extraction of key terms from the articles. 2) Construction of a term co-occurrence graph (in the scientific publications), 3) identifying densely connected subgraphs in this term co-occurrence graph and 4) inter-temporal analysis of dense subgraphs. The result of the analysis is represented in the form of phylomemetic lattices (which are analogous to phylogenetic trees that are used in biology, for representing the evolution natural species). While automatic phylomemetic structure reconstruction gives promising results, the scientist studying the evolution of science would like to interact with the tools and influence the construction algorithms.

The thesis should develop techniques that can enable the interactive construction of phylomemetic structures. Through the interaction the scientists can add or precise pieces of information in order to reduce the uncertainties present at the various stages of the reconstruction procedure.

The thesis will focus on some of the following aspects.

- Developing a model of phylomemetic structure as a (structured) knowledge extraction
- Enriching the extraction model with quality metrics
- We would like to develop algorithms that can support scientists exploring the graph (lattice). This requires data exploration techniques [8,9], as the phylomemetic structure is rather large in practice.

¹ <http://www-druid.irisa.fr/>

² <http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start>

- Provenance. As provenance questions can be important in the reconstruction process, our model should also deal with provenance information [10].
- Developing a workflow model of phylomemetic structure maintenance that can update parts of the network, in particular in the case of quality problems.

Competences

The PhD candidate should hold a master or equivalent degree in computer science. He or she should also have the following competences.

- Fluent in English (written, spoken)
- Good knowledge of data mining and knowledge extraction techniques
- Algorithmic and programming skills
- Ideally, experience with large-scale data management techniques

Foreign applications are welcome. French language skills are useful, but not mandatory.

Application

The PhD candidate should send by mail to zoltan.miklos@irisa.fr the following documents:

- Complete curriculum vitae
- Motivation letter
- Copy of grades at master level, and the master thesis (if available)
- Two references

Proposed starting date: **1st September 2017.**

References

- [1] Chavalarias, D. and Cointet, J-P. P. 2013. Phylomemetic patterns in science —the rise and fall of scientific fields. PloS one 8, 2, e54847.
- [2] Sun, X., Kaur, J., Milojevic, S., Flammini, A., Menczer, F. Social Dynamics of Science.
- [3] Takaffoli, M., Sangi, F. Fagnan, J., Zaiane, O. R. Community Evolution Mining in Dynamic Social Networks. Procedia – Social and Behavioral Science 22 (2011) 49-58
- [4] Berger-Wolf, T. Y., Saia, J. A Framework for Analysis of Dynamic Social Networks. KDD'2006, pp. 523-528.
- [5] Palla, G., Barabasi, A.-L., Vicsek, T. Quantifying social group evolution. Nature 446, pp. 664-667.
- [6] Asur, S., Parthasarathy, S. A viewpoint-based approach for interaction graph analysis. KDD'2009.
- [7] Malliaros, F. D., Megalooikonomou V., Faloutsos, C. Estimating robustness in large social graphs. Knowledge and Information Systems. 45(3), December, 2015, pp. 645-678
- [8] Ryen W. White, Resa A. Roth. Exploratory Search – Beyond the Query-Response Paradigm. Morgan & Claypool, 1999.
- [9] Peter T. Wood. Query languages for graph databases. SIGMOD Rec. 41(1) : 5060 (2012).
- [10] S. B. Davidson and J. Freire. Provenance and scientific workflows: Challenges and opportunities. In SIGMOD, 2008.