EPIQUE: A Graph Data Model and Query Language for Exploring the Evolution of Science

Ke Li, Hubert Naacke, Bernd Amann

EPIQUE: A Graph Data Model and Query Language for Exploring the Evolution of Science

Ke Li
LIP6, CNRS, Sorbonne Université
Paris, France
ke.li@lip6.fr

Hubert Naacke
LIP6, CNRS, Sorbonne Université
Paris, France
hubert.naacke@lip6.fr

Bernd Amann
LIP6, CNRS, Sorbonne Université
Paris, France
bernd.amann@lip6.fr

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; • **Information systems** → *Temporal data*; Data mining.

KEYWORDS

Topic Modeling, LDA, Science Evolution, Big data

Introduction. There is an increasing demand for practical tools to explore the evolution of scientific research published in bibliographic archives such as the Web of Science (WoS), ISTEX, arXiv or PubMed. The study of science evolution can help *philosophers and historians* of science [3] to test their theories with data, *researchers* to position their work in its scientific context, *industry* to evaluate the potential for innovation and technological transfer, *librarians* to classify scientific documents, etc. Revealing meaningful evolution patterns from document archives has many other applications and can be extended to synthesize narratives from datasets across multiple domains, including news stories, research papers, legal cases and works of literature [5].

In the interdisciplinary ANR EPIQUE project[1], we adopt the cognitive view of scientific evolution which assumes that the evolution only depends on the textual document contents (title, abstract, main contents) [3]. Whereas this choice reduces the expressive power by excluding the *social view* taking account of co-authorship and citation graphs [2, 6], it also decreases the "social" bias and detects more easily possible interactions between scientific ideas and contributions, independently of any particular scientific community. Graph-based topic evolution analysis builds on topic evolution networks [1] which track complex temporal evolution dynamics by periodical topic discovery and similarity-based topic alignment. Figure 1 shows a snippet of a topic evolution graph extracted from the arXiv[2] corpus. The graph covers the periods between 2000 and 2006 decomposed into three overlapping time periods (3 year periods with one year overlap). Each topic is represented by a rectangle containing the top-10 topic terms obtained by an NLP document pre-processing step. *Emerging* terms are shown in green, *decaying* term boxes are colored in red, *stable* terms which exist both, in ancestor topics and in descendant topics, are in blue and *specific* terms which appear only in the current topic are in white. The thickness of the alignment edges reflects the similarity of the

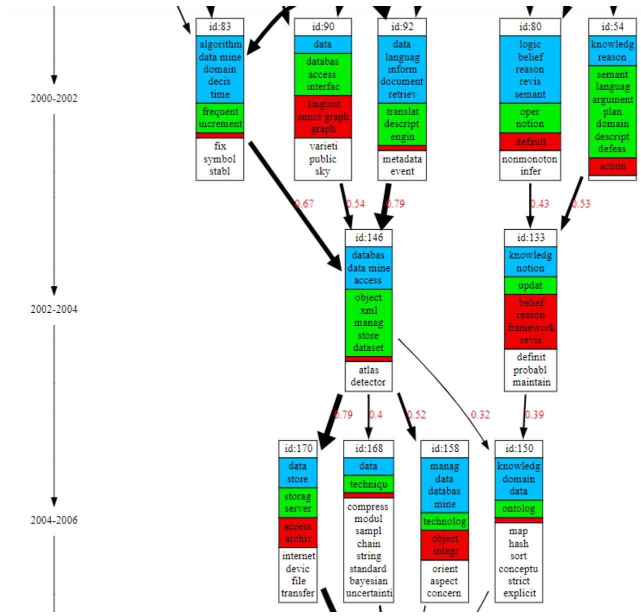

Figure 1: Pivot topics containing term "database" extracted from arXiv, green = emerging terms, blue = stable terms, red = decaying terms

connected topics. Several topics contain the term "database" and we can observe different evolution patterns. The topic evolution graph shows topics related to "data mining" (83), "data access interfaces" (90), "information retrieval" (92), "logics, semantics" (80) and "knowledge, reasoning" (54). The first three topics converge in 2002 – 2004 into a single topic on "object, xml, store, data mining" (146) which splits in the period of 2004 – 2006 into "storage servers" (170), "data mining and management" (158) and "knowledge and ontologies" (150).

Building and exploring topic evolution networks is still difficult and needs an important expertise in statistical text mining. A first challenge for domain experts is to correctly tune method specific hyper parameters with respect to a given dataset and an expected output. A second challenge concerns the visual exploration of large topic evolution networks. Whereas existing graph visualisation tools like Gephi [3] or Graphviz [4] can be used to generate high-quality visualisations, their use for exploring large graphs and identifying meaningful evolution patterns is difficult.

[1] This work was funded by French ANR-16-CE38-0002-01 project EPIQUE

[2] https://arxiv.org/

[3] https://gephi.org/
[4] https://www.graphviz.org/

Pivot Graph Model and Query Language. In this work we propose a data model for the visualisation and exploration of topic evolution networks representing the research progress in scientific document archives. Our model is independent of a particular topic extraction and alignment method and proposes a set of semantic and structural metrics for characterizing and filtering meaningful topic evolution patterns.

For identifying topic evolution patterns we decompose topic evolution graphs into subgraphs defined by a chosen topic t connected to other topics through alignment edges with some minimal similarity threshold β. Each couple (t, β) of some topic t and threshold β called a *pivot topic* and corresponds to a family of subgraphs $\mathcal{G}(t, \beta)$ called *pivot graphs*. We distinguish three particular pivot graphs denoted by (1) $\mathcal{G}^f(t, \beta)$, the maximal subgraph with all nodes that are reachable from t through paths with minimal edge weight β, (2) $\mathcal{G}^p(t, \beta)$, the maximal subgraph with all nodes that can reach t through paths with minimal edge weight β and their union (3) $\mathcal{G}^*(t, \beta) = \mathcal{G}^p(t, \beta) \cup \mathcal{G}^f(t, \beta)$.

The evolution of a topic t can then be characterized by the structure of its future $\mathcal{G}^f(t, \beta)$ and its past $\mathcal{G}^p(t, \beta)$ for different β-thresholds. The goal of our pivot graph model is to define a query language which allows users to filter topics according to some useful metrics concerning their evolution represented by their pivot graphs.

Our query language allows experts to filter pivot graphs according to some evolution pattern defined by the combination of graph evolution filters. For example query $Q1$ filters all pivot topics where the future has an average edge similarity (relative evolution degree) $Revol > 0.6$ and an average pivot topic similarity (pivot evolution degree) $Pevol > 0.5$, each future topic has two child topics in average ($Split$) and there exist future subtopics related to the pivot topic with a minimal distance of 5 periods ($Live$):

```
Q1: DB.Future.Revol(>=0.5).Pevol(>=0.6)
        .Split(>=2).Live(=5)
```

Observe that the user does not specify the β-threshold and the result contains for each topic t all its pivot topics (t, β) satisfying the filter.

Apart from these metric-based filters, our query language also allows users to define other multi-dimensional filtering criteria including topic labels and temporal conditions for the selection of pivot topics. For example, the following query *finds all topics with an emerging term "deep learning" where the past contains a path to a topic with the decaying term "big data"*:

```
Q2: DB.Emerge("deep_learning")
     .Past.Path(Decay("big_data"))
```

Finally, pivot topics and their associated metrics can be used for the structural and quantitative analysis of topic evolution graphs. For example Figure 2 shows the distribution of *future* pivot evolution graphs in arXiv with respect to their *split degree* and *convergence degree* . We can see that a low threshold $\beta = 0.2$ generates a large number of complex pivot topic graphs with high split and convergence degrees.

Implementation and Experimentation. The long version of this article includes a more detailed description of the underlying algorithms and other important aspects concerning quality issues like

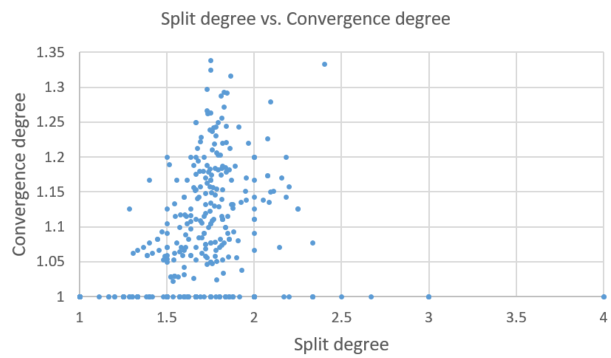

Figure 2: β = 0.2, #T = 50, #Pivot = 477, #Isolated = 23

topic diversity. The workflow also has been implemented on top of Apache Spark and we have have conducted several experiments on four real-world scientific archives covering 20 years of scientific publications including 1.15 million scientific articles extracted from arXiv and 1 million documents extracted from Wiley's Web Of Science.

Conclusion. In this article we propose a generic evolution network computation and visualization framework which combines a high-level data model with big data technology for extracting and exploring topic evolution networks. The graph model relies on the notion of *pivot topic graphs*, which describe the contents and the evolution dynamics of topics at different levels of detail. The model also includes a number of high-level semantic metrics which enable domain experts to specify meaningful topic evolution patterns (queries) for exploring large topic evolution networks. This framework has been completely implemented on top of Apache Spark using LDA and cosine similarity for topic extraction and topic alignment. The user can express complex evolution pattern queries to obtain the relevant pivot topic graphs. A first prototype [4] is used to extract complex evolution patterns for different scientific domains as part of the EPIQUE project and in collaboration with philosophers of science. As future work we intend to optimize the computation of pivot topic evolution graphs and exploit the LDA document-topic matrix for enriching the analysis.

REFERENCES

[1] David Chavalarias and Jean-Philippe Philippe Cointet. 2013. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one* 8, 2 (2013), e54847.

[2] Eugene Garfield. 1955. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122, 3159 (July 1955), 108–111.

[3] Thomas S. Kuhn, Otto Neurath, and Thomas Samuel Kuhn. 1994. *The Structure of scientific revolutions* (2nd ed., enlarged ed.). Number ed.-in-chief: Otto Neurath ; Vol. 2 No. 2 in International encyclopedia of unified science Foundations of the unity of science. Chicago Univ. Press, Chicago, Ill. OCLC: 258260085.

[4] Ke Li, Hubert Naacke, and Bernd Amann. 2020. EPIQUE: Extracting Meaningful Science Evolution Patterns from Large Document Archives (Demonstration). In *Int'l Conf. on Extending Database Technology (EDBT)*. Copenhagen, Denmark.

[5] Dafna Shahaf, Carlos Guestrin, Eric Horvitz, and Jure Leskovec. 2015. Information Cartography. *Commun. ACM* 58, 11 (2015), 62–73.

[6] Xiaoling Sun, Jasleen Kaur, Staša Milojević, Alessandro Flammini, and Filippo Menczer. 2013. Social Dynamics of Science. *Scientific Reports* 3 (Jan. 2013), 1069. https://doi.org/10.1038/srep01069