TrAnET: Tracking and Analyzing the Evolution of Topics in Information Networks

Livio Bioglio, Ruggero G. Pensa, and Valentina Rho

Dept. of Computer Science, University of Turin, Italy
{livio.bioglio,ruggero.pensa,valentina.rho}@unito.it

Abstract. This paper presents a system for tracking and analyzing the evolution and transformation of topics in an information network. The system consists of four main modules for pre-processing, adaptive topic modeling, network creation and temporal network analysis. The core module is built upon an adaptive topic modeling algorithm adopting a sliding time window technique that enables the discovery of groundbreaking ideas as those topics that evolve rapidly in the network.

Keywords: information diffusion · topic modeling · citation networks

1 Introduction

Information diffusion is an important and widely-studied topic in computational social science and network analytics due to its applications to social media/network analysis, viral marketing campaigns, influence maximization and prediction. An information diffusion process takes place when some nodes (e.g., customers, social profiles, scientific authors) influence some of their neighbors in the network which, in their turn, influence some of their respective neighbors. The definition of "influence" depends on the application. In mouth-to-mouth viral campaign, a user who bought a product at time t influence their neighbors if they buy the same product at time $t + \delta$. In bibliographic networks, author a influences author b when a and b are connected by some relationship (e.g., collaboration, co-authorship, citation) and either b cites one of the papers published by author a, or author b publish in the same topic as author a [2].

In this paper we propose a system for topic diffusion analysis based on adaptive and scalable Latent Dirichlet Annotation (LDA [1]) that uses a different notion of influence: for a given topic x, author a influences author b when b publish at time $t+\delta$ a paper that cites some papers covering topic x and authored by a at time t. Moreover, our focus is on topic evolution rather than on ranking authors, such as in [5]. Our system, in fact, enables the discovery of groundbreaking topics and ideas, which are defined as topics that evolve rapidly in the network. According to our definition, the most interesting topics are those that influence many new research topics, thus stimulating new research ideas. By setting different diffusion model parameters, our system enables the flexible analysis of topic evolution and the identification of the most influential authors. The salient

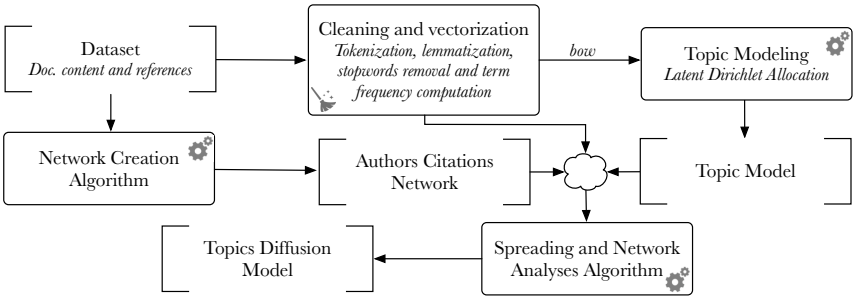


Fig. 1: A graphic overview of the overall processing and analysis pipeline.

features of our system, with respect to other state-of-the-art methods, are: (1) its ability to track the evolution and transformation of topics in time; and (2) its flexibility, enabling multiple types of online and offline analyses.

2 System Description

The architecture of the system is presented in Fig. 1. As an input, it takes a corpus consisting of any type of document (including scientific papers, patents, news articles) with explicit references to other previously published documents. First, the documents are pre-processed with NLP techniques that perform tokenization, lemmatization, stopwords removal and term frequency computation in order to prepare the corpus for the topic modeling module. This module adopts a scalable and robust topic modeling library [3] that enables the extraction of an adaptive set of to topics. Thanks to this module, it is possible to assign multiple weighted topics to a document published at time $t + \delta$ according to a topic model computed at the previous instant t. Moreover, the topic model can be adapted efficiently to newly inserted documents without recomputing it from scratch. A network creation module is used to extract the bibliographic network from the original corpus. Finally, the evolution of topic is tracked on the bibliographic network by a network analysis module that enables the visualization of several temporal characteristics of topic evolution, and the detection of the most interesting topics according to the evolution speed.

To perform topic evolution analysis, the spreading model considers several adjustable parameters. For each analysis task we consider: a time scale $[t_0, t_n]$ defining the overall time interval of the analysis; a time window of size δ and an overlap $\gamma < \delta$ defining a set of time intervals $\{\Delta T_0, \ldots, \Delta T_N\}$ s.t. $\forall i\ T_i = [t_0 + i(\delta - \gamma), t_0 + i(\delta - \gamma) + \delta)$; a set of K topics $\{\tau_1, \ldots, \tau_K\}$ (K being a user-given parameter). Given a topic τ_x, users in the network are activated at time ΔT_0 if they publish a paper covering topic τ_x during ΔT_0. Users are activated at time ΔT_i ($i > 0$) if they cite any paper that contributed to the activation of the users at time ΔT_{i-1}. A paper p is said to cover a topic τ_x if LDA has assigned τ_x to paper p with a weight greater than a user-specified threshold.

The whole process is driven within an interactive Jupyter notebook[1]. All modules are implemented in Python. All data are stored in a MongoDB[2] database server. The system runs on Windows, Linux and Mac OS X operating systems using a standard computing platform (e.g., any multi-core Intel Core iX CPU, and 8 GB RAM) and does not require any high-performance GPU architecture.

3 Demonstration

Dataset. The dataset used for TrAnET demonstration is a subset of the scientific papers citations network. This dataset is created by automatically merging two datasets originally extracted through ArnetMiner[4]: the DBLP and ACM citation networks[3]. The demonstration focuses on papers published from 2000 to 2014 within a set of preselected venues, for a total of about 155,000 papers.

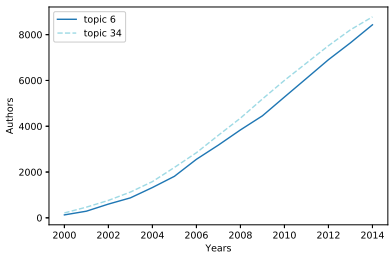
Text processing and topic extraction. The input data given to the topic extraction module is obtained as the result of the cleaning and vectorization process performed on the concatenation of paper title and abstract, as described in the previous section. In particular, the cleaning module ignores terms that appears only once in the dataset and in more than 80% of the documents. The topic extraction is performed on the whole dataset using Latent Dirichlet Allocation, searching for 50 topics. The topic model is then used to assign a weighted list of topics to each paper in the dataset. In our demonstration, we consider only topic assignments with weight greater than 0.2.

Example of topic evolution. To explain how our tool works, the analysis on two representative topics (namely, topics 6 and 34) is shown here: their keywords, sized according to their weight within the topic, are described in Fig. 2a and 2b, respectively. These topics have been chosen because they are assigned to a comparable number of papers (4498 for topic 6 and 6079 for topic 34) and authors (8430 for topic 6 and 8776 for topic 34). Moreover, they exhibit a very similar publication trend. According to Fig. 2c, which shows the cumulative number of authors that have published for the first time a paper on each topic in each year, the two trends are almost indistinguishable. This result (similar to what can be computed by [2]) shows that these topics have a similar diffusion trend in the bibliographic network. However, there is a strong difference in the evolution speed, as shown in Fig. 2d. Topic 34 (information retrieval) evolves more rapidly than topic 6 (clustering). This behavior can be explained by the increasing research efforts in the first field, driven by search engine and social media applications, as well as by Semantic Web technologies. Clustering, in contrast, appears as an evergreen albeit not particularly evolving research field in the time frame considered here. In this experiment, we used $K = 50$, $\delta = 4$ and

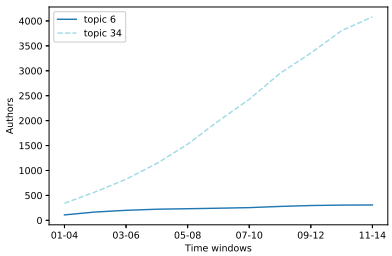
[1] `https://jupyter.org/`
[2] `https://www.mongodb.com/`
[3] `https://aminer.org/citation`

(a) Topic-6

(b) Topic-34



(c) Diffusion on network



(d) Speed of evolution

Fig. 2: Diffusion and word clouds of the selected topics.

$\gamma = 3$. By tuning the three parameters suitably, different outcomes will be shown during the demonstration.
The source code and the dataset of the demonstration are available online[4].

Acknowledgments. This work is partially funded by project MIMOSA (MultI-Modal Ontology-driven query system for the heterogeneous data of a SmArtcity, "Progetto di Ateneo Torino_call2014_L2_157", 2015-17).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Gui, H., Sun, Y., Han, J., Brova, G.: Modeling topic diffusion in multi-relational bibliographic information networks. In: Proceedings of CIKM 2014. pp. 649–658. ACM (2014)
3. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50 (2010)
4. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: KDD'08. pp. 990–998 (2008)
5. Xiong, C., Power, R., Callan, J.: Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of WWW 2017. pp. 1271–1279. ACM (2017)

[4] `https://github.com/rupensa/tranet`