DIGITAL LIBRARIES AND WEB INFORMATION SYSTEMS

5981P TEXT MINING PROJECT

# Topic and Trend Detection in Scientific Papers

*Submitted by:*
Monika Govindwar-82086
Prathmesh Halgekar-82721
Divyaben Hirpara-81956

*Supervised by:*
Prof. Dr. Siegfried Handschuh
Jelena Mitrovic

March 18, 2018

**Abstract**

Topic identification and trend detection is a new and challenging problem in text mining.The goal of this project is to identify academic topics and to detect trends in research using text mining techniques. Automatic identification of semantic content of documents has become increasingly important due to its effectiveness in many tasks including information retrieval, information filtering and organization of documents collections in digital libraries [**?**].

# Contents

# 1 Introduction

"What are the these **scientific papers** about ?"

As volume and diversity of scientific resources is growing at a rapid pace. Hence, topic identification from any scientific paper, trend detection and analysis have become much more important issues due to their application in many fields and the extensive growth of the number of documents in various domains. Usually, data mining tries to discover information hidden in scientific literature, which is not accessible by simple statistical techniques [**?**] whereas, text mining techniques area is significant subset of data mining that aims to extract knowledge from unstructured or semi-structured textual data and has widespread applications in analysing and processing textual documents [**?**]. Hence, we are using text mining techniques to identify topic and detect research trends in design research.

The knowledge of research topic or rather the early awareness of the emergence of a specific research topic would benefit anybody involved in the research environment. Imagine that we are researchers and looking for topics that have recently attracted much interest and utility in a particular domain. A manual review of all available articles in this domain would be so time-consuming as to be virtually impossible. In this situation, the automatic detection of emerging research trends can help researchers quickly understand the occurrence and the tendency of a scientific topic [**?**]. In returns, it will be more helpful for academic publishers and editors to exploit this knowledge and offer the most up to date and interesting contents.

# 2 Existing work and Motivation

Recently, several ETD (Emerging Trend Detection) models have been proposed [**?**], in which the ETD process can be viewed in three phases: topic representation, identification, and verification. The ETD central notion is usually represented by a set of temporal features in the topic

representation phase. These features are then extracted from document databases using text-processing methods in the feature extraction phase. After that, in topic verification these features are monitored over time and the topic is classified using interest and utility functions [**?**].

# 3 Applications or Technology used

The various tools and technologies are used to accomplish this project. Those are briefly discussed below.

- NLTK: The Natural Language ToolKit (NLTK) is a Python library for computational linguistics. NLTK includes a great number of common natural language processing tools including a tokenizer, a part of speech (POS) tagger, a stemmer, a lemmatizer which were extensively used in our project. In addition to these tools, NLTK has built in many common corpora including the Brown Corpus and WordNet.

- Scikit-learn: Sklearn is a machine-learning library for Python which provides simple and efficient tools for data analysis and data mining. It is designed to interoperate with the Python numerical and scientific libraries like NumPy, SciPy, and matplotlib. It features various classification, regression and clustering algorithms including k-means.

- NMF: Non-negative matrix factorization (NMF) in machine learning is unsupervised learning model where a matrix V is factorized into two matrices W and H, with the condition that all three matrices have no negative elements. NMF is widely applicable in most real world cases where V can't have negative values. General applications of nmf include: Topic recovery like Probabilistic Latent Semantic Analysis and Clustering like K-means. For clustering words, NMF is utilized in our project.

- NumPy: NumPy is a scientific and numerical computing extension used to operate on arrays in Python programming language. It supports for calculations with multi dimensional arrays.

# 4 System Architecture

# 5 System Requirements

# 6 Results

# 7 Conclusion

# 8 Future Enhancements

# References


1. Minh-Hoang Le, Tu-Bao Ho, Yoshiteru Nakamori, *Detecting Emerging Trends from Scientific Corpora*, Inter- national Journal of Knowledge and Systems Sciences, 2005.

2. Concept hierarchy `http://slideplayer.com/slide/8228328/`.

3. Francesco Osborne, Enrico Motta, Paul Mulholland, *Exploring Scholarly Data with Rexplore*,In The Semantic Web-ISWC 2013 (pp. 460-477)Springer Berlin Heidelberg. (2013).

4. G.Salton, C.S. Yang, "On the specification of term values in automatic indexing", Journal of Documentation, Vol. 29, pp. 351-372, 1973 `http://www.emeraldinsight.com/doi/pdfplus/10.1108/eb026562`.

5. Francesco Osborne, Angelo Salatino, Aliaksandr Birukou, Enrico Motta *Smart Topic Miner: Supporting Springer Nature Editors with Semantic Web Technologies*, In International Semantic Web Conference 2016 (pp. 383-399). Springer. (2016)

6. `https://en.wikipedia.org/wiki/Natural-language-processing`

7. Angelo A. Salatino, Francesco Osborne, Enrico Motta *How are topics born? Understanding the research dynamics preceding the emergence of new areas* https://peerj.com/articles/cs-119/

8. http://www.wsdm-conference.org/2016/slides/jie-tang-aminer.pdf