# Topic and Trend detection in Scientific Papers

Prathmesh Halgekar−82721 , Monika Govindwar−82086 , Divyaben Hirpara−81956

Prof. Dr. Siegfried Handschuh, Jelena Mitrovic
Chair of Digital Libraries and Web Information Systems
Text Mining Project

## Introduction

It has always been important certainly for academic publishers to annotate scholarly products with the appropriate research topics and keywords to facilitate the marketing process, to support digital libraries, academic search engines which in turn helps the research and academic communities. The knowledge of or rather the early awareness of the emergence of a specific research topic would significantly benefit anybody involved in the research environment. For e.g. academic publishers and editors can exploit this knowledge and offer the most up to date and interesting contents.

In this project we propose to:

1. Identify topic/topics in Scientific papers and

2. Detecting trends of a particular topic.

# 1 Topic Identification

## 1.1: Topic Representation

An important step in Topic Identification is topic representation. From existing sources [1] topics can be represented as a concept hierarchy as shown below:

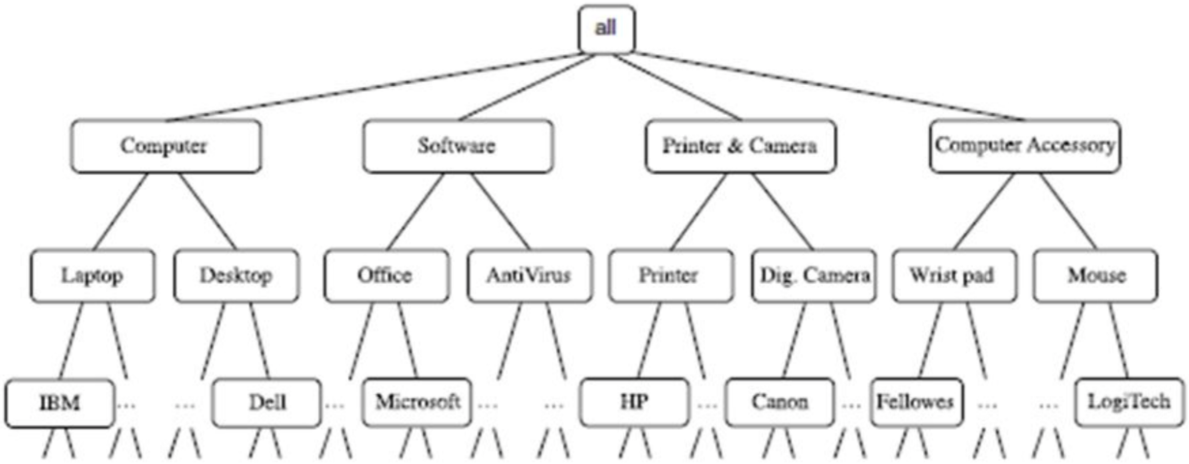

Figure 1: Concept Hierarchy [2]

Considering the above hierarchy and mapping it to Computer then the root node would be Computer and its child would be Laptops, Super computers, notebooks and etc. And even the child will further have its child as its various components.

The other approach would be to rely on an OWL ontology, which characterizes research areas and their relationships. This ontology is automatically populated and periodically updated by the Klink algorithm as mentioned in [3]

## 1.2: Topic Identification Process

This is an important part of automatic text processing techniques, such as information retrieval, text categorization, text summarization, etc. In this process we consider the Hybrid methods which combine the advantage of both statistical and knowledge-based methods to improve the robustness of identification process. As mentioned in [1] and [4] we use the $tf * idf$ measure to extract keywords from a document. Each keyword is then mapped into topics either in the concept hierarchy or in the OWL ontology using KLINK algorithm. The relevance of the document $d_j$ to the topic $t_i$ is computed as:

$$r(i,j) = \frac{Count(t_i)}{\sum_{t_j \in T} Count(t_j)} \tag{1}$$

where Count$(t_i)$ is the number of times the topic $t_i$ is counted.

# 2 Trend Detection

Let $S$ be a subset of $T$ where $t_i$ is a topic in $T$ but is not present in $S$. To compute the influence of topic $t_i$ on topics in $S$, first we will define $P(S)$ as the probability of topic in $S$ and $P(\bar{S})$ as the probability of topic not being in $S$ mentioned in an article. If the total number of articles is $n$ and the number of articles mentioning any topic in $S$ is $m$ then,

$$P(S) = \frac{m}{n} \tag{2}$$

and

$$P(\bar{S}) = \frac{n-m}{n} \tag{3}$$

The entropy of the occurrence of any topic in $S$ is:

$$H(S) = -P(S)logP(S) - P(\bar{S})logP(\bar{S}) \tag{4}$$

The entropy of the occurrence of any topic in $S$ under the condition that the occurrence of topic $t_i$ is known,can be calculated as follows:

$$H(\frac{S}{t_i}) = -\sum_{x=S,\bar{S}} \sum_{y=t_i,\bar{t_i}} P(x,y)logP(\frac{x}{y}) \tag{5}$$

Now, we will consider the mutual information:

$$I(S,t) = H(S) - H(\frac{S}{t}) \tag{6}$$

reflects the reduction in uncertainty about $S$ when the occurrence of $t_i$ is known.The greater $I(S,t_i)$ is, the more influence $t_i$ has on other topics in $S$.

# 3 Target Section

There are following applications or technologies important while working on this project.

(a) Smart Topic Miner [5]

(b) Rexplore [3]

(c) Emerging Trend Detection model [1]

# 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