http://www.jaist.ac.jp/iskss/

International Journal of
Knowledge and Systems Sciences
Vol. 2, No. 2, June 2005

Detecting Emerging Trends from Scientific Corpora

Minh-Hoang Le, Tu-Bao Ho, Yoshiteru Nakamori

School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Abstract

Emerging trend detection is a new challenge and an attractive topic in text mining. Our research goal was to construct a model to detect emerging trends in a set of scientific articles; the resulting model is richer in topic representation and more appropriate for evaluating emerging trends than existing models. To achieve this end, we associated each topic with many features extracted from scientific articles and constructed two measures for ranking interest and utility. Based on the information commonly provided in scientific papers, our method can adapt to different kinds of scientific corpora and also can be efficiently modified to adapt it to user needs. We also built a prototype system to test the model and the evaluations show that our model promises to achieve significant results in emerging trend detection.

Keywords: Emerging trend detection, Topic identification, Citation mining, Text mining

1. Introduction

Emerging trend detection (ETD) is a new and challenging problem in text mining. ETD is commonly defined as "detecting topic areas which are growing in interest and utility over time" [1].

A very significant task for ETD is to find emerging research trends in a collection of scientific articles. Imagine that we are researchers, 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, and thus they can, for example, find the most recent, related topics in their research domain.

Recently, several ETD models have been proposed [2-4], in which the ETD process can be viewed in three phases: topic representation, identification, and verification. Each topic – 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 [1]. Most work on ETD is centered around these three main tasks because the effectiveness of an ETD model completely depends on how appropriately a topic is represented in computers, how well the features associated with a topic are extracted from the documents, and how reasonably the interest and utility functions are constructed.

While many models have been proposed, most are still poor in representing research topics [3, 5] and inappropriate for determining and ranking interest and utility [4, 6]. There are two main reasons why these ETD models do not appear to be robust when applied to scientific corpora. First, many features can be extracted from scientific articles but may not be available in other textual data, meaning that these features cannot be integrated into general ETD models. The second limitation lies on the interpretation of interest and utility measures for evaluating research topics. This process is somewhat subjective and requires complex computations when analyzing the features associated with each topic.

The main objective of our work was to build an ETD model for scientific corpora. We have made three main contributions: (1) A rich representation scheme for topics using specific features of research articles, (2) Methods for extracting these features from documents, and (3) The interest and utility measures for evaluating emerging trends.

In the rest of this paper, Section 2 describes the structure of our model including methods for representing, identifying, and verifying topics. The experimental evaluations are given in Section 3. Our conclusion and future work will be discussed in the last section.

2. The Model for Emerging Trend Detection in Scientific Corpora

We propose the following emerging trend detection model:

$$M = \{D, T, f, g, CE\}$$

where

$D=\{d_j\}$: A set of scientific articles

$T=\{t_i\}$: A set of topics

$f(.)$: The measure of growth in interest

$g(.)$: The measure of growth in utility

C: The evaluator

E: The set of emerging trends

Our model has a set T consisting of topics to be evaluated. All topics in T are organized in a hierarchical thesaurus where each node is a topic and the hyponymy relationships between topics are considered as child-parent relationships in the concept hierarchy. The model takes a set of scientific articles D as the input and associates each topic with some temporal features extracted from D. After that, the evaluator C analyzes the time-series of features to evaluate the growth in interest and utility using two functions f and g and verifies whether or not each topic is an emerging trend. The output is a set of emerging trends selected by the model.

2.1 Topic Representation

Given a trial period of length Δ years, we represent each topic t_i in by a time series: $t_i = (t_i^1, t_i^2, \dots, t_i^{\Delta})$ where t_i^k represents t_i in the k^{th} year in the trial period. Each t_i^k is associated with 6 parameters:

t_i^k (1): determines how often the topic t_i is mentioned in the k^{th} year

t_i^k (2): The weight of citations in the k^{th} year to t_i, in which t_i is cited for referring to a theoretical basis, using methods or making comparison.

t_i^k (3): the number of citations in the k^{th} year to t_i

t_i^k (4): the influence of t_i on other topics in the k^{th} year

t_i^k (5): the weight of author reputations of t_i in the k^{th} year

t_i^k (6): the weight of sources (journals/proceedings) talking about t_i in the k^{th} year

In existing ETD models, a topic is often represented by n-grams, term frequencies, and term co-occurrences [2, 3] associated with date tags, author names, citations [4, 7], etc. In our model, we do not only consider individual topics, but also view each topic in its relation to others in order to examine its change in interest and utility over time. For this purpose, each topic is organized in the concept hierarchy and associated with many features extracted from scientific articles. This gives our model a richer representation scheme for topics so as to compute the growth in interest and utility more reasonably. Table 1 shows parameters associated with the topic "neural networks" extracted from three journals: Artificial Intelligence, ACM Transactions on Information Systems, and ACM Transactions on Modeling and Computer Simulation. The tendency of each parameter is displayed in Figure 1.

Table 1. Parameters associated with the topic "neural networks".

t_i=NNs	1998	1999	2000	2001	2002	2003
t_i^k (1)	0.06	0.10	0.08	0.10	0.09	0.06
t_i^k (2)	0.20	0.33	0.28	0.06	0.11	0.04
t_i^k (3)	0.41	0.40	0.50	0.12	0.07	0.32
t_i^k (4)	0.17	0.40	0.06	0.12	0.33	0.02
t_i^k (5)	0.65	0.55	0.13	0.24	0.67	0.11
t_i^k (6)	0.33	0.44	0.22	0.33	0.44	0.56

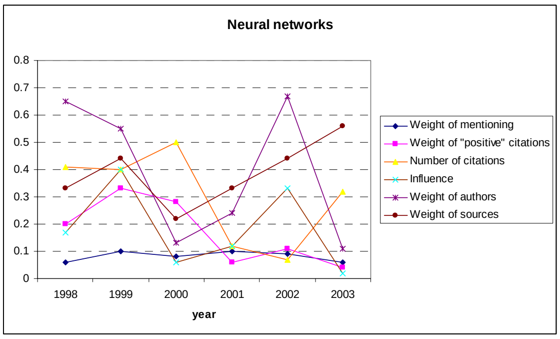

Fig. 1. The tendencies of parameters associated with the topic "Neural networks".

2.2 Topic Identification

To compute how often the topic t_i is mentioned in the k^{th} year, we have to measure the relevance of the topic t_i

to each article published in the k^{th} year. This problem, namely topic identification, is an important part of automatic text processing techniques, such as information retrieval, text categorization, text summarization, etc. Many methods of topic identification in these techniques can be divided into three groups: statistical methods, knowledge-based methods, and hybrid methods. Statistical methods [8, 9] infer the topic in the text from term frequencies, term locations, term co-occurrences, etc. without using external knowledge bases, whereas knowledge-based methods [10-12] rely on a syntactic/semantic parser, machine-readable dictionaries, etc. Hybrid methods [13, 14] combine the advantages of both statistical and knowledge-based methods to improve the robustness of the identification process.

In our ETD model, each topic in the concept hierarchy *T* is a set of synonymic words. The hierarchical structure is based on hyponymy relationships between topics. Figure 2 is an example of the topic “computer” and its sub-topics.

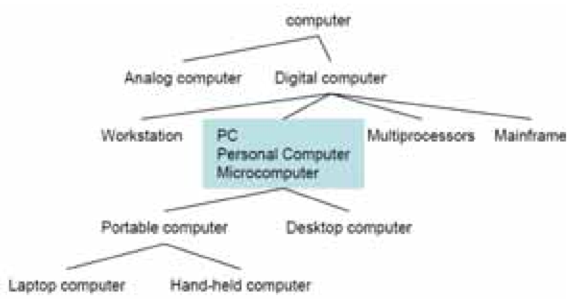

Fig. 2. The concept hierarchy.

The technique to compute the relevance of a topic $t_i \in T$ to an article $d_j \in D$ is follows: First, we use the tf∗idf measure [8] to extract keywords from d_j. Each keyword is then mapped into topics in the concept hierarchy *T*. Next, we scan the entire document to count how many times a topic is mentioned in d_j; note that whenever a topic is counted, its parent topics are also counted.

The relevance of the document d_j to the topic t_i is computed as:

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

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

To determine how often the topic t_i is mentioned in the k^{th} year, we sum up all relevances of documents published in the k^{th} year to t_i:

$$t_i^k(1) = \sum_{\text{year}(d_j)=k} r(i, j) \qquad (2)$$

Topic counting is an appropriate solution when a keyword matches several different topics. However, if a keyword has many different meanings, this method may be confusing when selecting the corresponding topic matched to the keyword. In this case, we have to analyze the context in which the keyword is placed and count the occurrences of all the words belonging to the topic and its ancestors/descendants in order to verify whether this keyword really matches this topic or not.

2.3 Citation Type Detection

Citations appear frequently in scientific articles and most digital libraries are now organized in the structure of citation indexes [15]. Examining the citations inside an article can reveal relationships between articles, draw attention to important corrections of published work, and identify significant improvements or criticisms of earlier work [16, 17]. However, to trace the development of a topic, we need a tool to identify the types of citation relationships that indicate the reasons for citation in a human-understandable way.

The purpose of identifying the reasons for citations (citation type detection - CTD) varies according to the main objective of each research project. Nanba and Okumura [18] use a heuristic sentence selection method and pre-defined cue phrases to classify citations into three categories to a system of automatic review articles. To extend the usage of linguistic patterns, Teufel [19] uses formulaic expressions, agent patterns, and semantic verb classes instead of cue phrases to determine the corresponding class for a sentence. Although both these works show the usefulness of linguistic patterns in citation type detection, the manual construction of linguistic patterns is obviously a rather time-consuming task. It also involves some conflicts that are difficult to resolve. For example, the method of Pham and Hoffmann [20] has to eliminate such conflicts and use human experts for adding rules that resolve them.

The available methods do not appear to be integrated into an ETD process because of two main limitations: first, their definitions of citation types are not appropriate for evaluating the interest and utility of topics; second, the manual construction of linguistic patterns must depend on the corpus, which makes the detection process inflexible when applied to other corpora. We want to create an appropriate definition of citation types which supports the detection of emerging trends by tracing the development of a topic and clarifying the relationship between articles. In addition, these citation types have to be automatically detected without any need for user-interaction or explicit knowledge about linguistic patterns, as were required in previous works.

In order to support researchers in tracing the development of a topic over time as well as clarify the relationship between articles, we classified citation types into the following six main categories (or classes), which are important for emerging trend detection:

- Type I: The paper is based on the cited work; this means that the citation shows other researchers' theories or methods as the theoretical basis for the current work.
- Type II: The paper is a part of the cited work
- Type III: The cited work supports this work
- Type IV: The paper points out problems or gaps in the cited work
- Type V: The cited work is compared with the current work
- Type VI: Other citations

Our method detects citation types using finite-state machines (FSMs). Two kinds of FSMs are considered: Hidden Markov Models (HMMs) [21] and Maximum-Entropy Markov Models (MEMMs) [22]. These are both stochastic models which have been successfully applied to many text-processing tasks such as language modeling, part-of-speech tagging, word-segmentation, named entity recognition, etc.

There are six FSMs corresponding to six classes that accept the set of English words including "\cite" as the set of observations. In the training phase, each FSM is given a number of training sentences for estimating its parameters. The Viterbi training algorithm [21] is used for training HMMs while Generalized Iterative Scaling algorithm is used to train MEMMs as described in the work of McCallum, Freitag and Pereira [21].

The synonymy and hyponymy relationships between words are represented in MEMMs using feature functions which have the form:

$$f_{(c,q)}(w,s) = \begin{cases} 1, \text{if } (s = q) \wedge (w \in c) \\ 0, \text{otherwise} \end{cases} \tag{3}$$

where *c* is a word concept, *w* is a word and *w*∈*c* means the concept *c* accepts the word *w* as its synonym or hyponym.

Representing synonymy and hyponymy relationships in HMMs is the post-processing task for the training phase: After estimating HMM parameters, we re-adjust the emission functions of HMMs:

$$\bar{b}_j(w) = \max_{o' \subseteq o} b_j(w') \tag{4}$$

where *w*′⊆*w* means the word *w*′ is a hyponym or synonym of the word *w*.

The citation type detection process can be described as follows: We are given six trained FSMs $\{\lambda_i\}_{i=1}^{6}$ and a citing area – a paragraph consisting of several sentences and the citation to be evaluated. For each sentence *O* in the citing area, we find the best state sequence Q_i^O corresponding to *O* in each machine λ_i and compute the likelihood $P^*(O|\lambda_i)=P(O,\ Q_i^O|\lambda_i)$ to measure how closely the sentence *O* matches the citation type *i*:

$$\begin{aligned} P^*(O \mid \lambda_i) &= \arg\max_{Q} P(O, Q \mid \lambda_i) \\ &= P(O, Q_i^{(O)} \mid \lambda) \end{aligned} \tag{5}$$

where $Q_i^{(O)}$ is the state sequence found by the Viterbi algorithm.

From equation (5), we compute

$$P(\lambda_i \mid O) = \frac{P^*(O \mid \lambda_i)}{\sum_{j=1}^{6} P^*(O \mid \lambda_j)} \tag{6}$$

as the probability of selecting model λ_i given the sentence *O*. The entropy of this probability distribution is:

$$H^{(O)} = -\sum_{i=1}^{6} P(\lambda_i \mid O) \cdot \log_2 P(\lambda_i \mid O) \tag{7}$$

As the entropy $H^{(O)}$ becomes larger, the chance of selecting the model corresponding to sentence *O* becomes more uncertain, and the role *O* plays in determining class label for the citing area becomes less important. Thus, we can weight each sentence *O* in the citing area by

$$Weight\ (O) = \frac{\log_2 6 - H^{(O)}}{\log_2 6} \qquad (0 \le Weight\ (O) \le 1) \tag{8}$$

Given a citing area *C* consisting of *m* sentences: O^1, O^2, …, O^m, the likelihood of *C* on category *i* is finally computed as:

$$L(C \mid i) = \frac{\sum_{j=1}^{m} Weight(O^j) \cdot P^*(O^j \mid \lambda_i)}{\sum_{i'=1}^{6} \sum_{j=1}^{m} Weight(O^j) \cdot P^*(O^j \mid \lambda_{i'})} \tag{9}$$

2.4 Calculating the Influence of a Topic

Let *S* be a subset of *T*, t_i is a topic in *T* but not belonging to *S*. We want to compute the influence of the topic t_i on topics in *S*. First, we define $P(S)$ $(P(\bar{S}))$ as the probability of any topic in *S* being (not being) 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 $P(S) = \frac{m}{n}$ and $P(\bar{S}) = \frac{n-m}{n}$.

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

$$H(S) = -P(S)\log P(S) - P(\bar{S})\log P(\bar{S}) \tag{10}$$

The entropy of the occurrence of any topic in *S* under the condition that the occurrence of topic t_i is known – $H(S|t_i)$ – can be calculated as follows:

$$H(S \mid t_i) = -\sum_{x=S,\bar{S}} \sum_{y=t_i,\bar{t}_i} p(x,y)\log p(x \mid y) \tag{11}$$

Now, we consider the mutual information:

$$I(S;t_i) = H(S) - H(S \mid t_i) \tag{12}$$

which 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*.

Because the set of articles used to compute the occurrence of a topic changes over time, we normalize the influence of t_i to the interval [0, 1]:

$$t_i^k(4) = \frac{I(T\setminus(t_i);t_i)}{H(T\setminus(t_i))} \qquad (13)$$

2.5 Building the Interest and Utility Measures

In our ETD model, growth is independently evaluated for all six time-series $\{t_i^k(j)\}_k$, ($1\le j\le 6$) and integrated into interest and/or utility functions. In concrete terms, the growth in interest of each topic is evaluated using four time-series $\{t_i^k(1)\}_k$, $\{t_i^k(3)\}_k$, $\{t_i^k(5)\}_k$, and $\{t_i^k(6)\}_k$; similarly, the growth in utility of each topic is evaluated using $\{t_i^k(2)\}_k$, $\{t_i^k(4)\}_k$, $\{t_i^k(5)\}_k$, and $\{t_i^k(6)\}_k$.

We propose two methodologies to calculate the growth in value over time: The first method measures the speed and acceleration of growth at a specific point, which can be viewed as a local evaluation. The second method uses inference to predict the dependence of the value on time. This is used to make a global evaluation for the tendency of the time-series in a given period.

To measure the speed and acceleration of a time-series, we first interpolate a time series $s=(s_1,s_2,\ldots,s_\Delta)$ by a continuous, smoothing function

$$\begin{aligned} & f[1,\Delta]\to R \\ & \text{st}: f(i)=s_i,(1\le i\le\Delta) \end{aligned} \qquad (14)$$

The level of growth in interest and utility of each topic at a specific point in time can be evaluated using these two measures. We can also classify topics in different ways according to their interest and utility: a topic growing fast in both interest and utility with high speed and high acceleration can be considered an emerging trend; a topic growing fast in interest but having small utility may be a new attractive research topic, and so on.

To make a global evaluation for the tendency of the a time-series we first consider each pair (*time*, *value*) as a data point, then use regression analysis to predict the dependence of values on the time. The simplest way is to apply linear regression on all data points and use the slope co-efficient of the regression equation to evaluate the global tendency of the time-series.

3. Experiments

We have developed a prototype system to test our model. The database contains 9000 full-length articles from our university library, originally in .PDF format. WordNet was used to detect synonymy and hyponymy relationships between words. We designed two experiments: the first experiment evaluated the effectiveness of the topic identification method and the second evaluated the accuracy of the citation type detection method. Evaluations of interest/utility functions and the entire ETD process are ongoing.

3.1 Evaluation of the Topic Identification Task

In this experiment, we performed tf å idf ranking and selected 1000 topics into the concept hierarchy *T*. For each article, we identified topics from the full text and from its abstract to compute three counts:

- *hits*: number of topics that are identified from full papers and also identified from their abstracts.
- *mistakes*: number of topics that are identified from full papers but are not identified from their abstracts.
- *misses*: number of topics that are not identified from full papers but are identified from their abstracts.

We then borrowed two measures from Information Retrieval:

- *Recall*: hits/(hits + misses)
- *Precision*: hits/(hits + mistakes)

The closer these two measures are to unity, the better the algorithm's performance. We randomly selected 100 papers for testing and achieved values of 0.52 and 0.58 in Recall and Precision respectively. When we added the keywords provided by the authors to the set of keywords extracted using tf∗idf, the accuracy was much improved: 0.82 in Recall and 0.87 in Precision.

3.2 Evaluation of citation type detection task

This experiment evaluated whether our method achieves higher accuracy compared to Nanba and Okumura's method when running under the same conditions. The data set provided by Nanba and Okumura in [18] consists of 282 citing areas for training and 100 citing areas for testing. We used the same definition of citation types as theirs: B, C and O and selected training sentences according to their sentence selection strategy. Table 2 shows the accuracy of Nanba and Okumura's method compared to our method.

Table 2. The accuracies of Nanba and Okumura's method, HMMs, and MEMMs.

	Nanba	HMMs	MEMMs
Type C	75.0%	87.5%	87.5%
Type B	78.1%	78.1%	81.3%
Type O	88.5%	92.5%	96.1%

Running under the same conditions, our method using HMMs and MEMMs based on concept-representation achieved higher accuracy than Nanba's method. Although the set of cue phrases is well designed for this dataset, Nanba's method still has the problem of synonymy and hyponymy, which is why our method using concept-representation can result in higher accuracy.

4. Conclusion

We have constructed a model for emerging trend detection in scientific databases in which we have developed computational methods for all model components, most of which are implemented in the prototype system.

Our model has a richer representation scheme for topics. Because the temporal features used for representing topics are based the information commonly provided in scientific papers, our model can adapt to different kinds of scientific corpora and also can be efficiently modified according to the needs of users.

In our experiments, the methods for topic identification and citation type detections achieved impressive results compared to other works. It is worth noting that these methods do not require user-interaction and their flexibility allows them to be extended.

Finally, the construction of interest and utility measures is a significant contribution of our work. By evaluating the growth in interest and utility separately, we can also classify emerging trends by different criteria as well as clarify the development of research topics in the published literature.

Acknowledgements

It is with genuine gratitude that we thank the Monbukagakusho for the scholarship that gave us the chance to work and study in Japan. We also wish to thank the COE project "Technology Creation Based on Knowledge Science" for funding our research.

References

[1] A. Kontostathis, L. Galitsky, W.M. Pottenger, S. Roy, D.J. Phelps, "A survey of emerging trend detection in textual data mining", in M. Berry (Ed.), *A Comprehensive Survey of Text Mining*, Chapter 9. Springer-Verlag, 2003.

[2] W.M. Pottenger, T.H. Yang, "Detecting emerging concepts in textual data mining", *Computational information retrieval*, pp. 89-105, 2001.

[3] R. Swan, J. Allan, "Automatic generation of overview timelines", in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49-56, ACM Press, New York, NY, USA, 2000.

[4] D.R. Gevry, *Detection of Emerging Trends: Automation of Domain Expert Practices*, 2002.

[5] J. Allan, R. Papka, V. Lavrenko, "On-line new event detection and tracking", in *Research and Development in Information Retrieval*, pp. 37-45, 1998.

[6] S. Havre, B. Hetzler, L. Nowell, "Themeriver: Visualizing theme changes over time", in *INFOVIS '00: Proceedings of the IEEE Symposium on Information Visualization 2000*, pp. 115, 2000.

[7] K. Rajaraman, A.H. Tan, "Topic detection, tracking and trend analysis using self-organizing neural networks", in *Proceedings of the Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01)*, 2001.

[8] G. Salton, C.S. Yang, "On the specification of term values in automatic indexing", *Journal of Documentation*, Vol. 29, pp. 351-372, 1973.

[9] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391-407, 1990.

[10] G. DeJong, "An overview of the frump system", in W.G. Lehnert, M.H. Ringle (Eds.), *Strategies for Natural Language Processing*, pp. 149-176, 1982.

[11] W.G. Lehnert, "Plot units: A narrative summarization strategy", in W.G. Lehnert, M.H. Ringle (Eds.), *Strategies for Natural Language Processing*, pp. 375-414, 1982.

[12] E. Riloff, W. Lehnert, "Information extraction as a basis for high-precision text classification", *ACM Transactions on Information Systems*, Vol. 12, No. 3, pp. 296-333, 1994.

[13] E.D. Liddy, S.H. Myaeng, "Dr-link's linguistic-conceptual approach to document detection", in *TREC*, pp. 113-130, 1992.

[14] M.A. Hearst, "Context and Structure in Automated Full-text Information Access", *PhD thesis*, University of California at Berkeley, 1994.

[15] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents", *Journal of the American Society of Information Science*, Vol. 24, pp. 265-269, 1973.

[16] S. Lawrence, C.L. Giles, K. Bollacker, "Digital libraries and autonomous citation indexing", *IEEE Computer*, Vol. 32, No. 6, pp. 67-71, 1999.

[17] R.N. Kostoff, J.A. del Rio, J.A. Humenik, E.O.

Garcia, A.M. Ramirez, “Citation mining: integrating text mining and bibliometrics for research user profiling”, *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 13, pp. 1148-1156, 2001.

[18] H. Nanba, M. Okumura, “Towards multi-paper summarization using reference information”, in *Proceedings of 16th International Joint Conference on Artificial Intelligence – IJCAI’99*, pp. 926-931, 1999.

[19] S. Teufel, “Argumentative zoning: Information extraction from scientific text”, *PhD thesis*, University of Edinburgh, 1999.

[20] S.B. Pham, A.G. Hoffmann, “A new approach for scientific citation classification using cue phrases”, in *Australian Conference on Artificial Intelligence*, pp. 759-771, 2003.

[21] L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition”, in *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[22] A. McCallum, D. Freitag, F. Pereira, “Maximum entropy Markov models for information extraction and segmentation”, in *Proceedings of the 17th International Conference on Machine Learning*, pp. 591–598, 2000.