IMPROVED METHOD FOR PATTERN DISCOVERY IN TEXT MINING

Bharate Laxman[1], D.Sujatha[2]

[1]Student, [2]Assistant professor, Department of CSE, ATRI, Andhra Pradesh, India
laxmanbharate@gmail.com**, **sujatha_dandu@gmail.com

Abstract

Digital data in the form of text documents is rapidly growing. Analyzing such data manually is a tedious task. Data mining techniques have been around to analyze such data and bring about interesting patterns. Many existing methods are based on term-based approaches that can't deal with synonymy and polysemy. Moreover they lack the ability in using and updating the discovered patterns. Zhong et al. proposed an effective pattern discovery technique. It discovers patterns and then computes specificities of patterns for evaluating term weights as per their distribution in the discovered patterns. It also takes care of updating patterns that exhibit ambiguity which is a feature known as pattern evolution. In this paper we implemented that technique and also built a prototype application to test the efficiency of the technique. The empirical results revealed that the solution is very useful in text mining domain.

Keywords – *Text mining, pattern discovery, text classification, pattern evolving*

---***---

1. INTRODUCTION

Knowledge discovery has become an indispensable phenomenon in recent years due to the rapid increase in digital data. They have attracted lot of attention in academic and scientific circles. Many applications in the real world need such mining of data in order to discover trends or patterns. These trends or patterns lead to business intelligence (BI). Such BI helps in taking well informed decisions. Many data mining techniques came into existence in the past ten years. They include closed pattern mining, maximum pattern mining, sequential pattern mining, item set mining, and association rule mining. These techniques are developed for data mining algorithms. They are capable of producing huge number of patterns. However, how to use those patterns and how to update them in future is the area that needs some more research. Especially in the field of text mining, patterns are discord from text documents. It is a challenging job to use those patterns and also update them. Earlier term based methods are provided by Information Retrieval (IR) techniques. The term based methods are classified into rough set models [1], SVM based models [2] and probability models [3]. All the term based methods suffer from problems such as synonymy and polysemy. When award has many meanings it is known as polysemy. When multiple words have similar meaning, it is called synonymy. Thus the discovered patterns with term based techniques have semantic meaning and answering the exact user query is difficult.

For this reason for many years people started believing that phrase-based techniques are better than that of term – based. However, the experiments in the field of data mining [4], [5], [6] have not been proved. The possible reasons include the phrases have less properties pertaining to statistics when compared with terms; frequency of occurrence is low; noisy and redundant phrases are more [6].

Though there are some drawbacks, the sequential patterns became promising alternatives to phrases [7], [8]. The reason for this is that sequential patterns avail required statistics like terms. Pattern Taxonomy Models (PTMs) [8], [9] came into existence to overcome the drawbacks of phrase-based mining approaches. Pattern based approaches became alternatives but much improvements are not made to make them more effective for text mining. With regard to effectiveness there are two issues. They are misinterpretation and low frequency. When patterns are less frequent, they can't be used for decision making. When the terms or patterns are misinterpreted, the result will not be reliable. Low frequency can't have required support. If the support is decreased, the results may not be useful for business decisions.

Over the last many years Information Retrieval (IR) is also used to have many techniques that used features of text documents. They are used to retrieve content from huge amount of documents based on the terms and their weights. The terms may have different weights based on the context as well. There might be semantic meanings that are to be considered in IR. Therefore it is not sufficient to only consider weights of terms for document analysis or evaluation. In this paper we implement a novel pattern discovery technique proposed by Zhong et al. [10]. It first computes specificities of the discovered patterns and then evaluates the weights of terms based on the distribution. Thus it is capable of avoiding misinterpretation problem. Negative training examples influence is also considered by this in order to avoid low-frequency problem. Moreover the ambiguous patterns are updated. This phenomenon is known as pattern evaluation.

Thus the proposed approach improves accuracy of the discovered patterns.

The remainder of this paper is organized as follows. Section II provides review of literature. Section III provides details of the proposed technique. Section IV presents implementation details. Section V provides experimental results while section VI concludes the paper.

2. PRIOR WORK

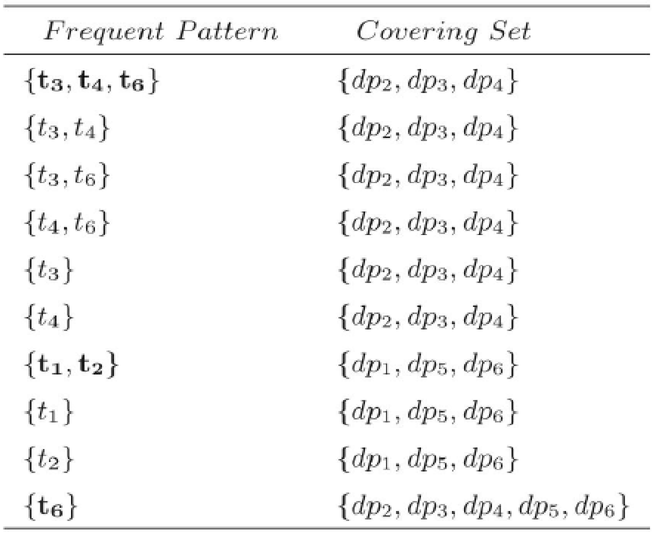
Textual documents are increasingly added to the World Wide Web and also the electronic databases of organizations. One of the representations which are well known is known as bag of words approach that makes use of keywords. Tf*idf weighting scheme is presented in [11] for representing text. In [12] entropy weighting and global IDF are used for text representation in addition to DFIDF. For the approach bag of words various schemes were developed for weighting [13], [14], [15]. The drawback in the bag of words is that choosing limited number of words is a problem thus it causes over fitting [6]. To reduce number of features other approaches came into existence. They include Odds ratio, Chi-Square, Mutual Information, and Information Gain [4], [6]. Though there are many representations, the choice of representation is based on the requirement, the rules of natural language [6].

Some researchers used phrases instead of words. Unigram and bigram combination is also used in the text categorization process. Phrase based approach is explored in [16]. Data mining techniques are also used as explored in [17]. There was no significant improvement in text mining when phrases are used. It suffered from lower frequency and misinterpretation problems [18]. Some insights were provided on ontology mining which is again term based [19], [20]. In [21] a technique known as pattern evolution was introduced. In data mining communities, pattern mining is extensively used for number of years. Algorithms such as GST [22], SLPMiner [23], SPADE [56] etc. are used for the purpose of data mining. However, finding interesting patterns is still open to anyone to research [25], [26]. Pattern mining is also used in text mining domain. Frequently found items is used text mining for various decisions making applications. Closed sequential patterns are also explored in text mining [9]. In [10] a model known as Pattern Taxonomy model is proposed in order to improve the discovered patterns in text mining. In [27] a two-stage model was developed. The two stages include pattern based methods and term based methods. For text mining "Natural Language Processing" concepts are used. Recently a new model known as concept-based model came into existence [28], [29]. Conceptual Ontology Graph is also explored in order to use semantic knowledge in the discovery of patterns. This model provides effective discrimination between meaningful terms and important terms. In this paper Pattern Taxonomy Model is used for text mining.

3. PATTERN TAXONOMY MODEL

The pattern taxonomy model described briefly here refers to Zhong et al. [10]. The PTM approach assumes that all text documents are converted to paragraphs. Therefore any given document is a set of paragraphs. By using "is a" relation it is possible to structure documents into taxonomy. Consider the following table.

Table 1: Frequent patterns and covering sets (excerpt from [10])



Frequent Pattern	*Covering Set*
$\{\mathbf{t_3}, \mathbf{t_4}, \mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{\mathbf{t_1}, \mathbf{t_2}\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{\mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_6\}$

As can be seen in table 1, frequent patterns are shows in the left column while the right column shows the documents in which these patterns exist. This is the based to structure pattern taxonomy. The constructed pattern taxonomy for the given values in table 1 is as shown in fig. 1.

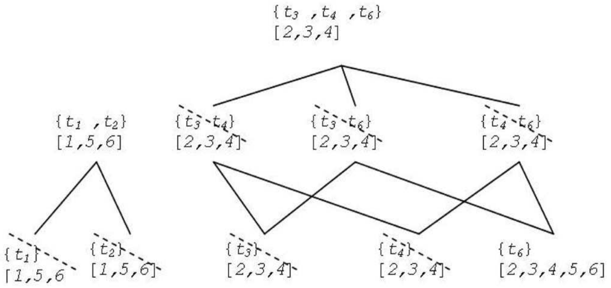
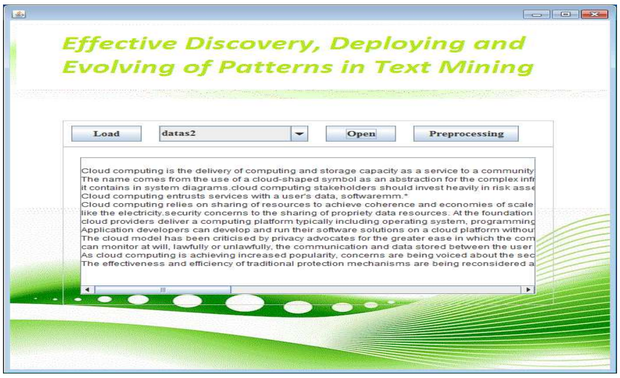


Fig. 1 – Pattern taxonomy (excerpt from [10])

As can be seen in fig. 1, there are many terms which are part of pattern taxonomy. This information is best used in text mining in order to produce closed patterns. The performance of text mining gets improved using this model. More details and deduced equations of this model are as explored in [10].

4. PROTOTYPE IMPLEMENTAITON

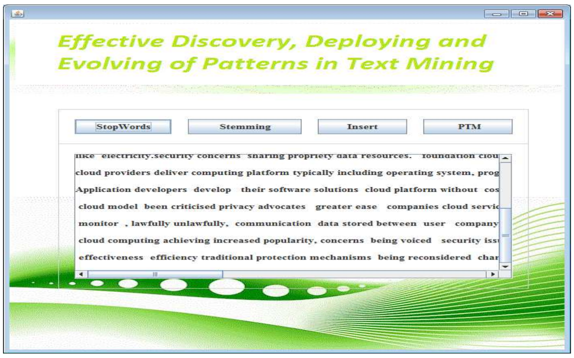
The pattern discovery technique proposed by Zhong et al. [10] has been implemented by us using Java programming language. The environment used for the implementation include a PC with 4GB RAM, Core 2 Dual processor. Operating system used is Windows and the IDE is Net Beans. Java SWING API is used to build GUI (Graphical User Interface). The main UI of the application is as shown in fig. 2.


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
Load
datas2
Open
Preprocessing

```
Cloud computing is the delivery of computing and storage capacity as a service to a community
The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infr
it contains in system diagrams.cloud computing stakeholders should invest heavily in risk asse
Cloud computing entrusts services with a user's data, softwaremm.*
Cloud computing relies on sharing of resources to achieve coherence and economies of scale
like the electricity.security concerns to the sharing of propriety data resources. At the foundation
cloud providers deliver a computing platform typically including operating system, programming
Application developers can develop and run their software solutions on a cloud platform withou
The cloud model has been criticised by privacy advocates for the greater ease in which the com
can monitor at will, lawfully or unlawfully, the communication and data stored between the user
As cloud computing is achieving increased popularity, concerns are being voiced about the sec
The effectiveness and efficiency of traditional protection mechanisms are being reconsidered a
```

Fig. 2 – The main UI of the prototype

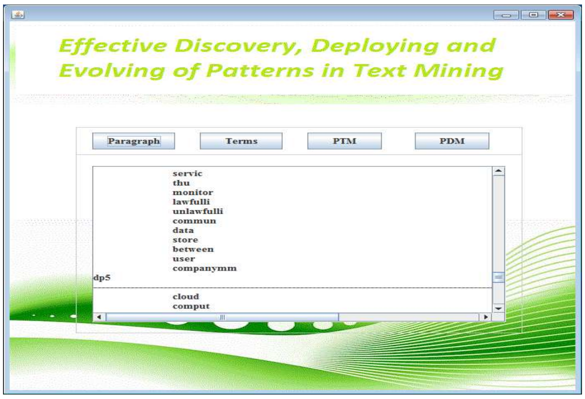
As seen in fig. 2 the application facilitates preprocessing before actual discovery of patterns. The selected dataset is shown in text area. Before proceeding further, the text needs to be preprocessed for operations like removal of stop words and stemming. On choosing preprocessing, the UI as shown in fig. 3 is rendered.


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
StopWords
Stemming
Insert
PTM

```
like  electricity.security concerns  sharing propriety data resources.   foundation clou
cloud providers deliver computing platform typically including operating system, prog
Application developers  develop   their software solutions  cloud platform without  cos
 cloud model  been criticised privacy advocates   greater ease   companies cloud servic
 monitor  , lawfully unlawfully,  communication  data stored between  user   company
 cloud computing achieving increased popularity, concerns  being voiced   security iss
 effectiveness  efficiency traditional protection mechanisms  being reconsidered  char
```

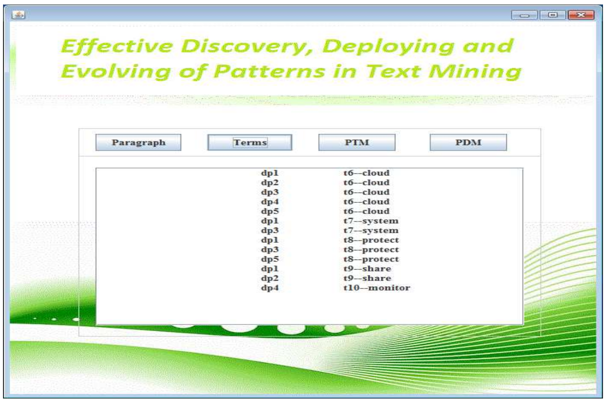
Fig. 3 – UI showing preprocessing operations

As can be seen in fig. 3, there is provision for stop words removal and stemming. These two are the fundamental pre-processing operations required before actually processing the text documents. The PIM button helps to build a pattern taxonomy model. The discovered patterns are shown in fig. 4.


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
Paragraph
Terms
PTM
PDM

```
            servic
            thu
            monitor
            lawfulli
            unlawfulli
            commun
            data
            store
            between
            user
            companymm
dp5
-----------------------------------------
            cloud
            comput
```

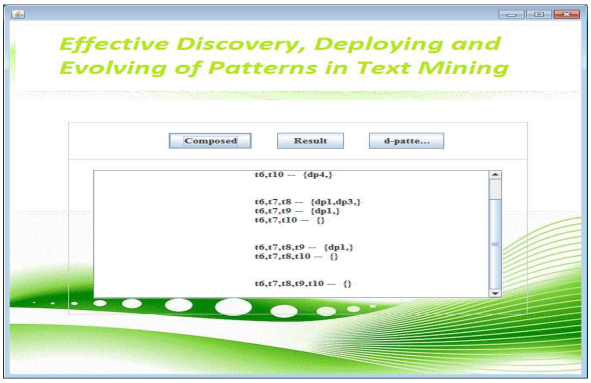
Fig. 4 – Discovered patterns


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
Paragraph
Terms
PTM
PDM

```
dp1        t6--cloud
dp2        t6--cloud
dp3        t6--cloud
dp4        t6--cloud
dp5        t6--cloud
dp1        t7--system
dp3        t7--system
dp1        t8--protect
dp3        t8--protect
dp5        t8--protect
dp1        t9--share
dp2        t9--share
dp4        t10--monitor
```

Fig 5 – Shows terms in the discovered patterns

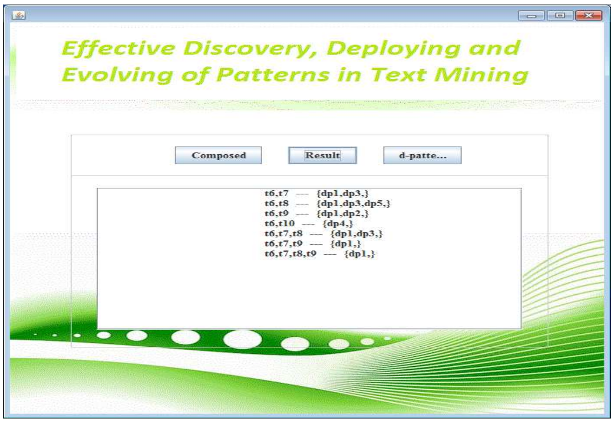
As can be seen in fig. 5, the terms for each discovered pattern are presented. The pattern1 has terms such as t6, t7, t8 and t9. This way the terms are shown for all discovered patterns. Groups of terms involved in different patterns are extracted and presented in fig. 6.


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
Composed
Result
d-patte...

```
t6,t10 --  {dp4,}


t6,t7,t8 --  {dp1,dp3,}
t6,t7,t9 --  {dp1,}
t6,t7,t10 --  {}


t6,t7,t8,t9 --  {dp1,}
t6,t7,t8,t10 --  {}


t6,t7,t8,t9,t10 --  {}
```

Fig. 6 – Distribution of Patterns and Terms

As can be seen in fig. 6, the discovered patterns distribution is shown. The terms are grouped according to the patterns to which they belong to. As can be observed in fig. 6, the results show the terms which do not belong to any patterns. Such terms are pruned and the results are presented in fig. 7.


Effective Discovery, Deploying and
Evolving of Patterns in Text Mining
Composed
Result
d-patte...

```
t6,t7 --- {dp1,dp3,}
t6,t8 --- {dp1,dp3,dp5,}
t6,t9 --- {dp1,dp2,}
t6,t10 --- {dp4,}
t6,t7,t8 --- {dp1,dp3,}
t6,t7,t9 --- {dp1,}
t6,t7,t8,t9 --- {dp1,}
```

Fig. 7 – Final results showing distribution of terms and patterns

As can be seen in fig. 7, the results reveal the distribution of terms and corresponding discovered patterns.

CONCLUSIONS

Data mining techniques have been around for long time. The techniques used to discover knowledge include sequential pattern mining, frequent item set mining, closed pattern mining and maximum pattern mining. These data mining techniques are not useful for text mining. This is due to lack of high specificity of discovered patterns. Not all frequent patterns discovered by mining algorithm are useful. Moreover then can be misinterpreted to make the problem worse. To overcome the problems of misinterpretation and low frequency, we proposed an effective pattern discovery. Pattern deploying and evolving are the two parts in the proposed technique. The empirical results revealed that the proposed technique is effective.

REFERENCES

[1] Y. Li, C. Zhang, and J.R. Swan, “An Information Filtering Model on the Web and Its Application in Jobagent,” Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[2]S. Robertson and I. Soboroff, “The Trec 2002 Filtering Track Report,” TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[4] D.D. Lewis, “Feature Selection and Feature Extraction for Text Categorization,” Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[5] S. Scott and S. Matwin, “Feature Engineering for Text Classification,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 379- 388, 1999.

[6] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[7] N. Jindal and B. Liu, “Identifying Comparative Sentences in Text Documents,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), pp. 244-251, 2006

[8] S.-T. Wu, Y. Li, and Y. Xu, “Deploying Approaches for Pattern Refinement in Text Mining,” Proc. IEEE Sixth Int’l Conf. Data Mining (ICDM ’06), pp. 1157-1161, 2006

[9] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, “Automatic Pattern- Taxonomy Extraction for Web Mining,” Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence (WI ’04), pp. 242-248, 2004.

[10] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[11] X. Li and B. Liu, “Learning to Classify Texts Using Positive and Unlabeled Data,” Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI ’03), pp. 587-594, 2003.

[12] S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991

[13] K. Aas and L. Eikvil, “Text Categorisation: A Survey,” Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[14] T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm withtfidf for Text Categorization,” Proc. 14th Int’l Conf. Machine Learning (ICML ’97), pp. 143-151, 1997

[15] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” Information Processing and Management: An Int’l J., vol. 24, no. 5, pp. 513-523, 1988.

[16] R. Sharma and S. Raman, “Phrase-Based Text Representation for Managing the Web Document,” Proc. Int’l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[17] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, “Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,” Proc. IEEE Int’lForum on Research and Technology Advances in Digital Libraries (ADL ’98), pp. 2-11, 1998

[18] D.D. Lewis, “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task,” Proc. 15th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’92), pp. 37-50, 1992.

[19] A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.

[20] C. Manning and H. Schu¨ tze, Foundations of Statistical Natural Language Processing.MIT Press, 1999

[21] Y. Li and N. Zhong, “Mining Ontology for Automatically Acquiring Web User Information Needs,” IEEE Trans.

Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.

[22] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003

[23] M. Seno and G. Karypis, "Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint," Proc. IEEE Second Int'l Conf. Data Mining (ICDM '02), pp. 418-425, 2002.

[24] M. Zaki, "Spade: An Efficient Algorithm for Mining Frequent Sequences," Machine Learning, vol. 40, pp. 31-60, 2001.

[25] Y. Li, W. Yang, and Y. Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 953-958, 2006.

[26] Y. Xu and Y. Li, "Generating Concise Association Rules," Proc. ACM 16th Conf. Information and Knowledge Management (CIKM '07), pp. 781-790, 2007

[27] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering," Proc. ACM 17th Conf. Information and Knowledge Management (CIKM '08), pp. 1023-1032, 2008.

[28] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006

[29] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007