International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 1 Issue 1 –AUGUST 2013

A Novel Way to Discover Patterns Adopting Effective Pattern Matching Approach for Knowledge Discovery Applications

Navuluri. Madhavilatha[#1], Bandla Srinivasa Rao[*2]

[#]*PG Scholar, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

[1] navuluri.madhavi@gmail.com

[*]*Associate Professor & HOD, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

***Abstract*-There are many data mining techniques have been proposed for the mining useful pattern in the text documents and the quality of the extracted features is the key issue to the text mining. The quality evidence for the existing text mining schemes due to the large of terms, phrases and noises. However, the data mining works with effectively, the update discover patterns having issues on the open research especially in the text mining as domain. The most existing text mining methods are adopted the term based approaches only, because they all suffer from the problems of polysemy and synonymy. However, the quality of the text mining may be not high because of lot of noise in text. For many years, the researchers research the hypothesis of pattern or phrase based approaches should perform better than the term based approaches, many experiments is conducted for the better results but experiments do not support the pattern based hypothesis. The failures of pattern based approaches due to the low frequency of occurrence, and include many redundant and noise phrases. To overcome the issues in the text mining, we introduce the new technique for getting the better performance. In this paper, we propose an innovative and effective pattern discovery technique for finding the relevant and interesting information in the text mining. To evaluate the proposed approach, we adopt the feature extraction method for information filtering (IF) and the experimental results conducted on Reuters Corpus Volume 1 and TREC topics confirm that the proposed approach could achieve excellent performance.**

***Index Terms*-Text mining, text classification, pattern mining, pattern evolving, information filtering, data mining, noise present, quality of mining.**

I. INTRODUCTION

There is a rapid growth in the computer and network technologies in recent years. In this technology, digital data also made available in the recent years and it show the fast growth in this field. This is due to the knowledge discovery and data mining have a great deal of attention with an imminent need for turning such data into useful information and knowledge. This type of technologies is easy to collect and stores in a large amount of unstructured or semi-structured texts such as webpage's, HTML/XML archives, emails, and text files. And these text data can be an idea with the large scale text databases, it becomes important to develop efficient tools to discover interesting knowledge from such text databases. There are many applications such as business management and market analysis; it can be benefits with knowledge and information extracted from a large amount of data. Data mining is therefore an essential step in the process of knowledge discovery in databases.

In the last ten years, the numbers of data mining techniques have been proposed for the different knowledge tasks. The proposed data mining techniques are closed pattern mining, frequent item set mining, sequential pattern mining, association rule mining and association rule mining. The purposes of the proposed mining techniques are to find the particular patterns within a reasonable and to adequate the time frame. . With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. In this paper, we mainly focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

Text mining works to find the interesting knowledge in text documents. It is one of the challenging issues to find the knowledge in text documents to help user to find what they want really. To solve the above issue, Information Retrieval (IR) provides many term based methods. Some of the term-based models are Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM), these are all based on the filtering models. . The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. Though, term based methods provides services and it suffer from the problems of polysemy and synonymy. Whereas polysemy means the single word has multiple meanings and synonymy means the multiple words have the same meanings.

However, the quality of the text mining may be not high because of lot of noise in text. In order to overcome the issues in the term based models, the researchers approach their idea to the hypothesis of pattern or phrase based models should perform better than the term based approaches, many experiments is conducted for the better results but experiments

do not support the pattern based hypothesis. The failures of pattern based approaches due to the low frequency of occurrence, and include many redundant and noise phrases. Sequential patterns used in data mining community have turned out to be a promising alternative to phrases [13], [50] because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantages of the phrase based models, we propose to the pattern mining-based approaches. This proposed model consists of the two patterns for the effectiveness in the text mining. The two patterns are: the concept of closed sequential patterns, and pruned non-closed patterns. These pattern mining-based approaches have shown certain extent improvements on the effectiveness. However, the paradox is that people think pattern-based approaches could be a significant alternative, but consequently less significant improvements are made for the effectiveness compared with term-based methods.

However the researchers introduce the pattern based models, it consists of two disadvantages Low frequency and Misinterpretation. A highly frequent pattern (normally a short pattern with large support) is usually a general pattern, or a specific pattern of low frequency. If we decrease the minimum support, a lot of noisy patterns would be discovered. Misinterpretation means the measures used in pattern mining turn out to be not suitable in using discovered patterns to answer what users want. The difficult problem hence is how to use discovered patterns to accurately evaluate the weights of knowledge in text documents. In the pattern based model has issues so the researchers IR has developed many mature techniques that terms were important features in the text documents. . However, many terms with larger weights are general terms because they can be frequently used in both relevant and irrelevant information.

In order to solve the problems in the text mining techniques we propose the new technique as an effective pattern discovery for text mining. In this technique, first calculates the discovered patterns and then evaluates term weights according to the distribution of terms. In this discovered pattern rather than the distribution in documents for solving the misinterpretation problem in the text mining. The influence of the distributed patterns from the negative training is to find the ambiguous patterns and then its try to solve the low frequency problems. In the proposed model, it can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents in the text mining. We conducted many experiments on the data collection, Reuters Corpus Volume1 (RCV1) and Text Retrieval Conference (TREC) filtering topics to prove the proposed models are effective and accuracy than the previous models in the text mining. The results show that the proposed technique outperforms up-to-date data mining-based methods, concept-based models and the state-of-the-art term based methods.

The rest of the paper is organized as follows. In Section II, we discuss about the related work of the text mining. In Section III formally introduces our proposed system of the paper. In Section IV we summarize about the algorithm used in the model. In Section V, we present the full simulation study of the proposed scheme. Finally, we conclude the paper and discuss future work in Section VI.

II. Related Works

In this section, we briefly discuss the works which is similar techniques as our approach but serve for different purposes.

Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvi, Tanya Clement, Ben Shneiderman, and Catherine Plaisant [57], This paper addresses the problem of making text mining results more comprehensible to humanities scholars, journalists, intelligence analysts, and other researchers, in order to support the analysis of text collections. Our system, Feature Lens, visualizes a text collection at several levels of granularity and enables users to explore interesting text patterns. The current implementation focuses on frequent item sets of n-grams, as they capture the repetition of exact or similar expressions in the collection. Users can find meaningful co-occurrences of text patterns by visualizing them within and across documents in the collection. This also permits users to identify the temporal evolution of usage such as increasing, decreasing or sudden appearance of text patterns. The interface could be used to explore other text features as well. Initial studies suggest that Feature Lens helped a literary scholar and 8 users generate new hypotheses and interesting insights using 2 text collections

Ah-Hwee Tan [53], Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values. Last count reveals that there are more than ten high-tech companies offering products for text mining. Has text mining evolved so rapidly to become a mature field? This article attempts to shed some lights to the question. We first present a text mining framework consisting of two components: Text refining that transforms unstructured text documents into an intermediate form; and knowledge distillation that deduces patterns or knowledge from the intermediate form. We then survey the state-of-the-art text mining products/applications and align them based on the text refining and knowledge distillation functions as well as the intermediate form that they adopt. In conclusion, we highlight the upcoming challenges of text mining and the opportunities it offers.

M. Rajman, and R. Besancon [17], In the general framework of knowledge discovery, Data Mining techniques are usually dedicated to information extraction from structured databases. Text Mining techniques, on the other hand, are dedicated to information extraction from

unstructured textual data and Natural Language Processing (NLP) can then be seen as an interesting tool for the enhancement of information extraction procedures. In this paper, we present two examples of Text Mining tasks, association extraction and prototypical document extraction, along with several related NLP techniques.

sMany types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In [21], the tf*idf weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in [9] and improves performance by an average of30 percent. Various weighting schemes for the bag of words representation approach were given in [1], [14], [38]. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid overfitting [41]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. Details of these selection functions were stated in [19], [41].

In [7], the combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase-based text representation for Web document management was also proposed in [44]. In [3], data mining techniques have been used for text analysis by extracting co-occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned in [18].

Term-based ontology mining methods also provided some thoughts for text representations. For example, hierarchical clustering [28], [29] was used to determine synonymy and hyponymy relations between keywords. Also, the pattern evolution technique was introduced in [25] in order to improve the performance of term-based ontology mining.

Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms [2], [31], [49], PrefixSpan [32], [53], FP-tree [10], [11], SPADE [56], SLPMiner [42], and GST [12] have been proposed. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [22], [24], [52]. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent item sets, co-occurring terms and multiple grams, for building up a representation with these new types of features. Nevertheless, the challenging issue is how to effectively deal with the large amount of discovered patterns.

Recently, a new concept-based model [45], [46] was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. This model included three components. The first component analyzed the semantic structure of sentences; the second component constructed a conceptual ontological graph (COG) to describe the sematic structures; and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between nonimportant terms and meaningful terms which describe a sentence meaning. Compared with the above methods, the concept-based model usually relies upon its employed NLP techniques.

III. PROPOSED WORK

In this paper, we mainly focus on the issues of an existing In order to solve the problems in the text mining techniques we propose the new technique as an effective pattern discovery for text mining. In this technique, first calculates the discovered patterns and then evaluates term weights according to the distribution of terms. In this discovered pattern rather than the distribution in documents for solving the misinterpretation problem in the text mining. The influence of the distributed patterns from the negative training is to find the ambiguous patterns and then its try to solve the low frequency problems. In the proposed model, it can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents in the text mining. We conducted many experiments on the data collection, Reuters Corpus Volume1 (RCV1) and Text Retrieval Conference (TREC) filtering topics to prove the proposed models are effective and accuracy than the previous models in the text mining. The results show that the proposed technique outperforms up-to-date data mining-based methods, concept-based models and the state-of-the-art term based methods.

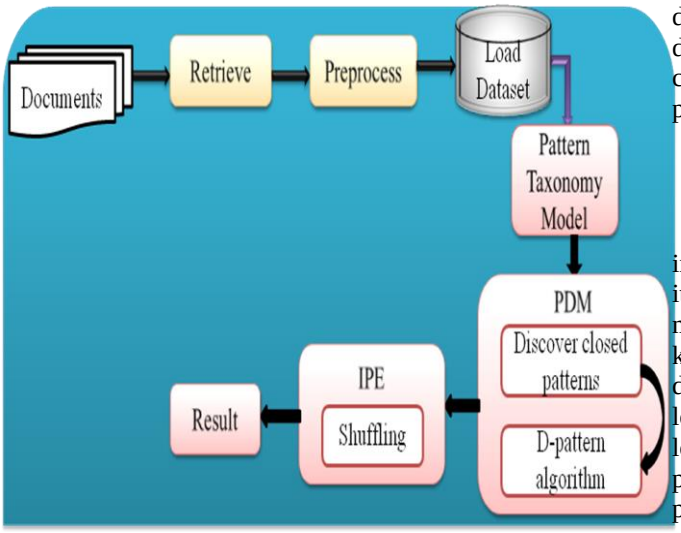

Figure 1: Our Proposed system

IV. Simulation Works/Results

We have simulated our system in Java. We implemented and tested with a system configuration on Intel Dual Core processor, Windows XP and using Netbeans 7.0. We have used the following modules in our implementation part. The details of each module for this system are as follows. We have implemented and tested with the 5 modules.

Loading document

In this module, to load the list of all documents. The user to retrieve one of the documents. This document is given to next process. That process is preprocessing.

Text Preprocessing

The retrieved document preprocessing is done in module. There are two types of process is done. Stop words removal text stemming. Stop words are words which are filtered out prior to, or after, processing of natural language data. Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

Pattern taxonomy process

In this module, the documents are split into paragraphs. Each paragraph is considered to be each document. In each document, the set of terms are extracted. The terms, which can be extracted from set of positive documents.

Pattern deploying

The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated.

Pattern evolving

In this module used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. If partial conflict offender contains in positive documents, the reshuffle process is applied.

V. Conclusion

There are many data mining techniques have been proposed in the last ten years includes association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. In this discovered knowledge (or patterns) in the field of text mining is having difficulties and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. In data mining, misinterpretations of patterns lead to the ineffective performance so researchers works for an an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

VI. References

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

[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 478-499, 1994.

[3] 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’l Forum on Research and Technology Advances in Digital Libraries (ADL ’98), pp. 2-11, 1998.

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

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, “Kernel Methods for Document Filtering,” TREC, trec.nist.gov/ pubs/trec11/papers/kermit.ps.gz, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, “Word-Sequence Kernels,” J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, “Statistical Phrases in Automated Text Categorization,” Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell’Informazione, 2000.

[8] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[9] 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.

[10] J. Han and K.C.-C. Chang, “Data Mining for Web Intelligence,” Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[11] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), pp. 1-12, 2000.

[12] 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.
[13] 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.
[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.
[15] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning (ICML '98),, pp. 137-142, 1998.
[16] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.
[17] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
[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] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
[20] D.D. Lewis, "Evaluating and Optimizing Automous Text Classification Systems," Proc. 18th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '95), pp. 246-254, 1995.
[21] 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.
[22] 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.
[23] 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.
[24] Y. Li and N. Zhong, "Interpretations of Association Rules by Granular Computing," Proc. IEEE Third Int'l Conf. Data Mining (ICDM '03), pp. 593-596, 2003.
[25] 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.
[26] 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.
[27] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.
[28] Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.
[29] Manning and H. Schu¨ tze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
[30] Moulinier, G. Raskinis, and J. Ganascia, "Text Categorization: A Symbolic Approach," Proc. Fifth Ann. Symp. Document Analysis and Information Retrieval (SDAIR), pp. 87-99, 1996.
[31] J.S. Park, M.S. Chen, and P.S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '95), pp. 175-186, 1995.
[32] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," Proc. 17th Int'l Conf. Data Eng. (ICDE '01), pp. 215-224, 2001.
[33] M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, 1980.
[34] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.
[35] S.E. Robertson, S. Walker, and M. Hancock-Beaulieu, "Experimentation as a Way of Life: Okapi at Trec," Information Processing and Management, vol. 36, no. 1, pp. 95-108, 2000.
[36] J. Rocchio, Relevance Feedback in Information Retrieval. chapter 14, Prentice-Hall, pp. 313-323, 1971.
[37] T. Rose, M. Stevenson, and M. Whitehead, "The Reuters Corpus Volume1—From Yesterday's News to Today's Language Resources," Proc. Third Int'l Conf. Language Resources and Evaluation, pp. 29-31, 2002.
[38] 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.
[39] M. Sassano, "Virtual Examples for Text Classification with Support Vector Machines," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '03), pp. 208-215, 2003.
[40] S. Scott and S. Matwin, "Feature Engineering for Text Classification," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 379-388, 1999.
[41] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
[42] 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.
[43] R.E. Shapire and Y. Singer, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, vol. 39, pp. 135-168, 2000.
[44] 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.
[45] 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.
[46] 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.
[47] K. Sparck Jones, S. Walker, and S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments—Part 1," Information Processing and Management, vol. 36, no. 6, pp. 779-808, 2000.
[48] K. Sparck Jones, S. Walker, and S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments—Part 2," Information Processing and Management, vol. 36, no. 6, pp. 809-840, 2000.
[49] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. 21th Int'l Conf. Very Large Data Bases (VLDB '95), pp. 407-419, 1995.
[50] 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.
[51] 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
[52] Y. Xu and Y. Li, "Generating Concise Association Rules," Proc. ACM 16th Conf. Information and Knowledge Management (CIKM '07), pp. 781-790, 2007.
[53] X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2003.
[54] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval, vol. 1, pp. 69-90, 1999.
[55] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99), pp. 42-49, 1999.
[56] M. Zaki, "Spade: An Efficient Algorithm for Mining Frequent Sequences," Machine Learning, vol. 40, pp. 31-60, 2001.
Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvi, Tanya Clement, Ben Shneiderman, and Catherine Plaisant-
" Discovering interesting usage patterns in text collections: Integrating text mining with visualization"