Journal of Theoretical and Applied Information Technology
31st August 2017. Vol.95. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SURVEY ON DEVELOPMENT OF PATTERN EVOLVING MODEL FOR DISCOVERY OF PATTERNS IN TEXT MINING USING DATA MINING TECHNIQUES

RAVINDRA CHANGALA, DR.D RAJESWARA RAO
CSE Department, K L University, Andhra Pradesh

ABSTRACT

As the data increasing, retrieving useful information and knowledge for the users is an open issue in the text mining domain in spite of having many existed data mining techniques. We focused on this by conducting a survey over existed techniques with our new method called pattern discovery models. We found that the existed model yields few drawbacks like low frequency problem, effective usage of discovered patterns, noisy data, polysemy and synonymy etc. Most people felt with the hypothesis that pattern based methods will perform better than the term based techniques. We got support to prove our model is better example for effective use of patterns in text mining than the existed.

Keywords: *Text Mining, Pattern Evolving, Pattern Deploying, Information Retrieval, Pattern Taxonomy Model, Data Mining Techniques.*

1. INTRODUCTION

As digital technology increased as well as e-data getting increased abundantly. The situation became very difficult for the user to get useful information from large amounts of data. Among those fields text mining one where the user can perform efficiently. In the last decade many such type of data mining techniques have been proposed but they suffered from frequency problem, effective usage of discovered patterns, noisy data, polysemy and synonymy etc. Hence we conducted a survey on these issues by comparing different models. As in the process of survey we known that pattern methods performed better than term based methods[1].

Structure of Data

Due to digital technology scope increased while data capacity becoming huge. The data may contain improper structured and not organized well. Hence getting useful patterns are difficult. In this view we first focused on the structure of the data. On the type of data is unstructured data in which documents having raw text may contains dates, numbers, facts in the documents is difficult to understand by the traditional programs.

Many techniques have been proposed to get useful patterns from unstructured data as data mining, natural language processing(NLP) and text analytics these may interpret the information and also text mining is an efficient technique. The Unstructured Information Management Architecture (UIMA) standard provided a common framework for processing this information to extract meaning and create structured data about the information.

The **unstructured** data information does not have a pre-defined data model or is not organized in a pre-defined manner. Examples of "unstructured data" may include books, journals, documents, metadata, healthrecords, audio, video , analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. Unstructured data is a generic label for describing data that is not contained in a database or some other type of data structure . Unstructured data can be textual or non-textual. Textual unstructured data is generated in media like email messages, PowerPoint presentations, Word documents, collaboration software and instant messages. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and Flash video files.

Structured data (Pre-defined and machine-readable, is locatable and usually has a relational 'data model' and usually is about real-world objects). Structured data refers to kinds of data with a high level of organization, such as information in a relational database. When information is highly structured and predictable, search engines can more easily organize and display it in creative ways. Structured data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

markup is a text-based organization of data that is included in a file and served from the web. It typically uses the schema.org vocabulary an open community effort to promote standard structured data in a variety of online applications. Structured data markup is most easily represented in JSON-LD format, which stands for JavaScript Object Notation for Linked Data. For the most part, structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations; whereas unstructured data is essentially the opposite.

Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.

2. TEXT MINING

Text mining may be named as text data mining or text analytics is the process of getting effective usable information from text data bases by devising patterns since text is structured and unstructured. The phrase text mining is the process of any system analyzes huge amount natural language text to produce useful information [1]. Finally text mining is about looking for patterns in text. Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want.

In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models [4], rough set models [23], BM25 and support vector machine (SVM) [34] based 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. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

Patterns in Text Mining

A pattern is a regular and intelligible form or sequence discernible in the way in which something happens or is done. For example the pattern “LIB” is given less weight age than the pattern “JDK” because of low confidence and support. It is difficult to evaluate weight of the term LIB since inadequate scope. This is one of the aspects of our work to give solution.

3. RELATED WORK & LITERATURE SURVEY

Digital technology leads to effective growth of the digital data, handling which is an issue. Hence the importance of knowledge discovery and data mining role increased to give useful information and knowledge to the user by using its techniques association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Because of this many domains benefited like market analysis, business management etc. All those concentrated generation of large number of patterns but failed how to utilize them effectively.

Regarding these issues we conducted a survey which made us confident to further precede our work efficiently. In the survey we found few issues like polysemy and synonymy, the idea of many people having that pattern based approaches perform better than the term based, but many experiments do not support this hypothesis. There are two fundamental issues regarding the effectiveness of pattern-based approaches as low frequency and misinterpretation. We also conducted immense survey on experiments done on the latest data collection, Reuters Corpus Volume 1 (RCV1) and Text Retrieval Conference (TREC) filtering topics, to evaluate the proposed technique. The survey 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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We also gone through the models such as Rocchio and probabilistic models rough set models [23], BM25 and support vector machine (SVM) [34] based filtering models. In our survey we identified that even term based model performing well but suffering from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning and failed to provide exact pattern to the user due to phrases have inferior statistical properties to terms, they have low frequency of occurrence and there are large numbers of redundant and noisy phrases among them. The alternate for these are effective pattern discovery model, uses the concepts of sequential patterns statistical properties like terms, pattern mining-based approaches having closed sequential patterns, and pruned nonclosed patterns.

Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modeling of hidden patterns. In this context the selection of characteristics and also the influence of domain knowledge and domain-specific procedures play an important role. Therefore, an adaptation of the known data mining algorithms to text data is usually necessary.

4. KNOWLEDGE DISCOVERY

Knowledge discovery is the process of nontrivial extraction of information from large databases, information that is implicitly present in the data, previously unknown and potentially useful for users. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

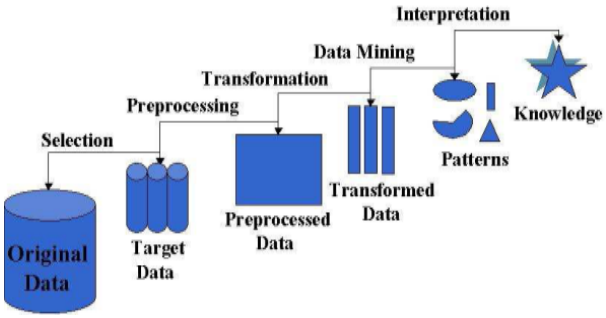


Fig 1. A Complete Knowledge Discovery Model

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).

Basic Definitions KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Here, data are a set of facts (for example, cases in a database), and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. It consists of an iterative sequence of the following steps:

Data cleaning (to remove noise and inconsistent data)
Data integration (where multiple data sources may be combined)
Data selection (where data relevant to the analysis task are retrieved from the database)
Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures).
Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. DATA MINING METHODS FOR TEXT

Data mining studies focused more on structured data but not unstructured data such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages. Most of the data now a days available in electronic form in the text databases as semistrucutred hence importance to the modeling and implementation of it became great deal while information retrieval techniques needed more. The existed retrieval techniques are not adequate since analyzing and extracting useful information from huge data bases. User approached different tools regarding text data analysis and information retrieval to have useful information.

Basic Measures for Text Retrieval: Precision and Recall

In this section we discuss about the basic measures for text retrieval. The user may get documents based on his query given to the system. Among those documents how many retrieved documents are relevant to his work is a question, can be measured by using two techniques precision and recall.

Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

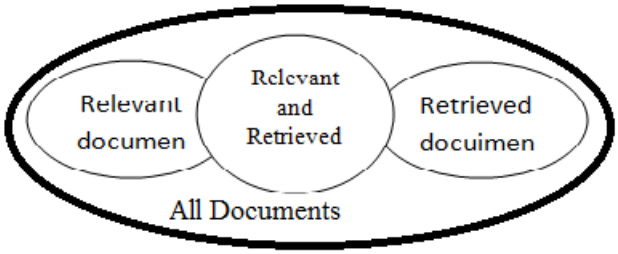
$$precision = \frac{|\{Relevant\} \cap \{Retrieved}{|\{Retrieved\}|}$$

Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision:

$$F\_score = \frac{recall \times precision}{(recall + precision)/2}.$$



Relationship between the set of relevant documents and the set of retrieved documents

6. TEXT RETRIEVAL METHODS

We have two types of information retrieval methods which are document selection problem or as a document ranking problem. In document selection methods, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the Boolean retrieval model. Document ranking methods use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods. . There are many different ranking methods based on a large spectrum of mathematical foundations, including algebra, logic, probability, and statistics.

The goal these methods are to approximate the *degree of relevance* of a document with a score computed based on information such as the frequency of words in the document and the whole collection. The (weighted) term-frequency matrix *TF*(*d*; *t*) measures the association of a term *t* with respect to the given document *d*: it is generally defined as 0 if the document does not contain the term, and nonzero otherwise.

$$TF(d,t) = \begin{cases} 0 & if\, freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & otherwise. \end{cases}$$

Besides the term frequency measure, there is another important measure, called inverse document frequency (IDF), that represents the scaling factor, or the importance, of a term *t*.

$$IDF(t) = \log \frac{1 + |d|}{|d_t|},$$

where *d* is the document collection, and *dt* is the set of documents containing term *t*. In a complete vector-space model, TF and IDF are combined together, which forms the TF-IDF measure:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t).$$

A term frequency matrix where each row represents a document vector, each column represents a term, and each entry registers *freq*(*di*; *t j*), the number of occurrences of term*t j* in document *di*.
Based on this table we can calculate the TF-IDF value of a term in a document. For example, for *t*6 in *d*4, we have

$$TF(d4; t6) = 1+\log(1+\log(15)) = 1.3377$$

$$IDF(t6) = \log(1+5/3)= 0.301$$

Therefore,

$$TF\text{-}IDF(d4; t6) = 1.3377x0.301 = 0:403$$

A representative metric is the cosine measure, defined as follows. Let *v*1 and *v*2 be two document vectors. Their cosine similarity is defined as $sim(v1;v2) = v1 \cdot v2/|v1| |v2|$

A term frequency matrix showing the frequency of terms per docume

document/term	t_1	t_2	t_3	t_4	t_5	t_6
d_1	0	4	10	8	0	5
d_2	5	19	7	16	0	0
d_3	15	0	0	4	9	0
d_4	22	3	12	0	5	15
d_5	0	7	0	9	2	4

There are several popular text retrieval indexing techniques, including inverted indices and signature files. An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: document table and term table, where document table consists of a set of document records, each containing two fields: doc id and posting list, where posting list is a list of terms (or pointers to terms) that occur in the document, sorted according to some relevance measure. term table consists of a set of term records, each containing two fields: term id and posting list, where posting list specifies a list of document identifiers in which the term appears.

7. TEXT INDEXING TECHNIQUES

There are several popular text retrieval indexing techniques, including *inverted indices* and *signature files*. An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: *document table* and *term table*, where *document table* consists of a set of document records, each containing two fields: *doc id* and *posting list*, where *posting list* is a list of terms (or pointers to terms) that occur in the document, sorted according to some relevance measure. *term table* consists of a set of term records, each containing two fields: *term id* and *posting list*, where *posting list* specifies a list of document identifiers in which the term appears.

Query Processing Techniques: Once an inverted index is created for a document collection, a retrieval system can answer a keyword query quickly by looking up which documents contain the query keywords.

Dimensionality Reduction for Text

However, for any nontrivial document database, the number of terms *T* and the number of documents *D* are usually quite large. Such high dimensionality leads to the problem of inefficient computation, since the resulting frequency table will have size *TxD*. Furthermore, the high dimensionality also leads to very sparse vectors and increases the difficulty in detecting and exploiting the relationships among terms (e.g., synonymy).
To overcome these problems, dimensionality reduction techniques such as latent semantic indexing, probabilistic latent semantic analysis, and locality preserving indexing can be used.

8. TEXT MINING APPROACHES

There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. In general, the major approaches, based on the kinds of data they take as input, are: (1) the keyword-based approach, where the input is a set of keywords or terms in the documents, (2) the tagging approach, where the input is a set of tags, and (3) the information-extraction approach, which inputs semantic information, such as events, facts, or entities uncovered by information extraction.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Keyword-Based Association Analysis

Such analysis collects sets of keywords or terms that occur frequently together and then finds the association or correlation relationships among them. Like most of the analyses in text databases, association analysis first preprocesses the text data by parsing, stemming, removing stop words, and so on, and then evokes association mining algorithms.

The problem of keyword association mining in document databases is thereby mapped to item association mining in transaction databases, where many interesting methods have been developed. Term recognition and term-level association mining enjoy two advantages in text analysis: (1) terms and phrases are automatically tagged so there is no need for human effort in tagging documents; and (2) the number of meaningless results is greatly reduced, as is the execution time of the mining algorithms. With such term and phrase recognition, term-level mining can be evoked to find associations among a set of detected terms and keywords. Therefore, based on user mining requirements, standard association mining or max-pattern mining algorithms may be evoked.

Document Classification Analysis

Automated document classification is an important text mining task because, with the existence of a tremendous number of on-line documents, it is tedious yet essential to be able to automatically organize such documents into classes to facilitate document retrieval and subsequent analysis. Document classification has been used in automated topic tagging (i.e., assigning labels to documents), topic directory construction, identification of the document writing styles (which may help narrow down the possible authors of anonymous documents), and classifying the purposes of hyperlinks associated with a set of documents.

Document Clustering Analysis: Document clustering is one of the most crucial techniques for organizing documents in an unsupervised manner.

9. TEXT MINING ALGORITHMS

We used the following text mining algorithms for our work.

Algorithm	Function
Naive Bayes	Classification
Generalized Linear Models	Classification, Regression
Support Vector Machine	Classification, Regression, Anomaly Detection
k-Means	Clustering
Non-Negative Matrix Factorization	Feature Extraction
Apriori	Association Rules
Minimum Descriptor Length	Attribute Importance

10. CONCLUSIONS

In this survey we had much knowledge about patterns, text mining and other techniques of text mining. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support. We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. The survey we studied existed about 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.

REFERENCES

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

[2] R. Agrawal and R. Srikant, “Fast Algorithms for MiningAssociation Rules in Large Databases,” Proc. 20th Int’l Conf. VeryLarge 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 PhraseExtraction in Digital Document Collections,” Proc. IEEE Int’l Forum on Research and Technology Advances in Digital Libraries(ADL ’98), pp. 2-11, 1998.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[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 Phrasesin Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazionedell'Informazione, 2000.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks," MachineLearning, vol. 20, no. 3, pp. 273-297, 1995.

[9] S.T. Dumais, "Improving the Retrieval of Information fromExternal Sources," Behavior Research Methods, Instruments, andComputers, 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 withoutCandidate Generation," Proc. ACM SIGMOD Int'l Conf. Managementof Data (SIGMOD '00), pp. 1-12, 2000.

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

[13] N. Jindal and B. Liu, "Identifying Comparative Sentences in TextDocuments," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research andDevelopment in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithmwith tfidf for Text Categorization," Proc. 14th Int'l Conf. MachineLearning (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 UsingSupport 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 Categorizationand Its Application to Text Retrieval," IEEE Trans. Knowledgeand Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[18] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representationson a Text Categorization Task," Proc. 15th Ann. Int'l ACMSIGIR Conf. Research and Development in Information Retrieval(SIGIR '92), pp. 37-50, 1992.

[19] D.D. Lewis, "Feature Selection and Feature Extraction for TextCategorization," Proc. Workshop Speech and Natural Language,pp. 212-217, 1992.

[20] D.D. Lewis, "Evaluating and Optimizing Automous Text ClassificationSystems," Proc. 18th Ann. Int'l ACM SIGIR Conf. Researchand Development in Information Retrieval (SIGIR '95), pp. 246-254,1995.

[21] X. Li and B. Liu, "Learning to Classify Texts Using Positive andUnlabeled 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 forRepresentations 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 Modelon the Web and Its Application in Jobagent," Knowledge-BasedSystems, vol. 13, no. 5, pp. 285-296, 2000.

[24] Y. Li and N. Zhong, "Interpretations of Association Rules byGranular Computing," Proc. IEEE Third Int'l Conf. Data Mining(ICDM '03), pp. 593-596, 2003.

[25] Y. Li and N. Zhong, "Mining Ontology for AutomaticallyAcquiring Web User Information Needs," IEEE Trans. Knowledgeand 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 TextMining 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. MachineLearning Research, vol. 2, pp. 419-444, 2002.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[28] A. Maedche, Ontology Learning for the Semantic Web. KluwerAcademic, 2003.

[29] C. Manning and H. Schu¨ tze, Foundations of Statistical NaturalLanguage Processing.MIT Press, 1999.

[30] I. Moulinier, G. Raskinis, and J. Ganascia, "Text Categorization: ASymbolic Approach," Proc. Fifth Ann. Symp. Document Analysis andInformation Retrieval (SDAIR), pp. 87-99, 1996.

[31] J.S. Park, M.S. Chen, and P.S. Yu, "An Effective Hash-BasedAlgorithm for Mining Association Rules," Proc. ACM SIGMODInt'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, andM. Hsu, "Prefixspan: Mining Sequential Patterns Efficiently byPrefix-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 TrackReport," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.ps.gz.

[35] S.E. Robertson, S. Walker, and M. Hancock-Beaulieu, "Experimentationas a Way of Life: Okapi at Trec," Information Processingand 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 CorpusVolume1—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 inAutomatic 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 withSupport Vector Machines," Proc. Conf. Empirical Methods in NaturalLanguage 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 FindingFrequent Sequential Patterns Using Length-Decreasing SupportConstraint," 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 Systemfor Text Categorization," Machine Learning, vol. 39, pp. 135-168,2000.

[44] R. Sharma and S. Raman, "Phrase-Based Text Representation forManaging the Web Document," Proc. Int'l Conf. InformationTechnology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[45] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text ClusteringUsing 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 forEnhancing Text Categorization," Proc. 13th Int'l Conf. KnowledgeDiscovery and Data Mining (KDD '07), pp. 629-637, 2007.

[47] K. Sparck Jones, S. Walker, and S.E. Robertson, "A ProbabilisticModel of Information Retrieval: Development and ComparativeExperiments—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 ProbabilisticModel of Information Retrieval: Development and ComparativeExperiments—Part 2," Information Processing and Management,vol. 36, no. 6, pp. 809-840, 2000.

[49] R. Srikant and R. Agrawal, "Mining Generalized AssociationRules," 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 PatternRefinement in Text Mining," Proc.