ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Relevant Feature Discovery from Text Documents Using Text Mining

Vaibhav Kudke[1], Akash Tapkir[2], Shubham Salunke[3], Yuvraj Kunjir[4], Prof. N.S. Shirsat[5]

BE Students, Department of Information Technology, PVG's COET, Pune, SavitribaiPhule Pune University Pune India[1,2,3,4]

Professor, Department of Information Technology, PVG's COET, Pune, SavitribaiPhule Pune University Pune India [5]

ABSTRACT: In this paper we introduce a method to select irrelevant documents for weighting features. We continued to develop the RFD model and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method. This paper presents an innovative model for relevance feature discovery. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms). It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns. Substantial experiments using this model on RCV1, TREC topics and Reuters-21578 show that the proposed model significantly outperforms both the state-of-the-art term-based methods and the pattern based methods.

KEYWORDS: Data mining feature selection, information retrieval, text classification.

I. INTRODUCTION

The objective of relevance feature discovery (RFD) is to find the useful features available in text documents, including both relevant and irrelevant ones, for describing text mining results. This is a particularly challenging task in modern information analysis, from both an empirical and theoretical perspective. This problem is also of central interest in many Web personalized applications, and has received attention from researchers in Data Mining, Machine Learning, and Information Retrieval and Web Intelligence communities. There are two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents. The first is the low-support problem. Given a topic, long patterns are usually more specific for the topic, but they usually appear in documents with low support or frequency. If the minimum support is decreased, a lot of noisy patterns can be discovered.Web search engines return arrangements of website pages sorted by the page's relevance to the user query. The issue with web search relevance ranking is to establish relevance of a page to a query [12]. These days, business web-page search engines combine hundreds of features to approximate relevance [13].Information Retrieval (IR) Systems are the associates of Web and search engines. These systems are designed to retrieve documents from digital collections e.g. library abstracts, corporate reports, news and so forth. Generally, IR relevance ranking algorithms are designed to obtain high recall on medium sized document collections using detailed user queries. Moreover, textual documents in these collections had practically no structure or hyperlinks [12]. A web search engine uses many methods of the standards and calculations of Information Retrieval Systems, however needed to adjust and stretch out them to fit their needs. Data mining techniques help user to find valuable information from a huge amount of text documents on the Web. Many text mining techniques have been developed in order to get the goal of retrieving useful information for users [12]. Most of them accept the term-based approach whereas the others choose the pattern-based technique to create a text representative for a set of documents. Information Retrieval has provided many efficient term-based techniques to solve this challenge [17]. The benefits of term-based technique include efficient computational performance. In the recent work, various data mining techniques have been proposed for feature

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

(e.g. term, pattern) discovery. These tasks include sequential pattern mining, frequent pattern mining and closed pattern mining. The synonymy and polysemy are the main issues associated with term-based methods [3], [9], and [11]. Polysemy implies same word has numerous meaning while synonymy implies a different word has the same meaning [3].Also pattern-based methods face low frequency and miss understanding problems [3]. A highly relevant pattern is usually a specific pattern of low frequency. Many noisy patterns are discovered, if we reduce the minimum support. The measures used in pattern mining (support and confidence) turn out to be not suitable to discover useful patterns which lead to miss understanding. In text document, the complicated task is how to use discovered patterns to precisely evaluate the weights of useful feature [3], [12].

1.1 Motivation

Search engine retrieve a list of thousands or millions of web pages based on user query. Manydata mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue,especially in the domain of text mining. Pattern based approaches have shown encouragingimprovements on effectiveness. However, two challenging issues have arisen when pattern mining techniques were introduced for IR systems. In the presence of these set backs, some studies adopted data mining to discover various patterns in text. Such patterns have the potential for text mining since they have predictive power, and allow to capture semantic relationships existing Department of Computer Engineering among terms in sentences, paragraphs, or even the whole document. Moreover, data mining has developed advanced methods for eliminating redundant patterns and noisy patterns. Motivated by the above problems researcher introduce an pattern discovery approach.

1.2 Objectives

2. Ranking the relevant document according to the topic specific term occurrence.
3. Categories the terms from the relevant documents into three categories: Positive, Negative, General categories.

II. LITERATURE SURVEY

These days web assets and its use is continuously increasing much over the time. Userneeds valuable data rapidly, while utilizing web. There are a large number of new documentsin web and user want efficient results while searching the web. There are some issues in Websearch [12], such as effective ranking and relevance, evaluation and information needs. The IRcommunity faces the challenge of managing a huge amount of hyperlinked data, but membersof this community can utilize modeling, document classification and categorization, user interfaces,and data visualization altering to accomplish their goals [12] [13]. Information Retrievalmodels are based on ranking algorithm, which is used in search engines to generate the rankedlist of documents [6]. A ranking algorithm sorting a set of documents according to their relevanceto a give query [8].Feature selection is the method of selecting a subset of relevant features for use in modelcreation. In text documents feature can be term, pattern, sentence. However, the traditionalfeature selection techniques are not efficient for selecting text features for solving the relevanceproblem because relevance is a single class problem [13]. The well-organized way of featureselection technique for relevance and techniques is based on a feature weighting function. Afeature weighting function indicates the amount of information represented by the feature occurrencesin a document and indicates the relevance of the feature.The term-based Information Retrieval models contain the Rocchio algorithm [13], [19],Probabilistic models, language model and Okapi BM25 [19]. In a language model, the keyconcept is the probabilities of word sequences which include both sentences and words. Theyare commonly approximated by n-gram models [13], like Unigram, Bigram and Trigram, forconsider term dependency. In the recent work important issue for feature selection in a textdocument is to identify format of the document. Text feature can be a single word or complexstructure. It comprises various complex structures such as n-grams, pattern and term.The objects can be phonemes, syllables, letters, words or base pairs according to the application.Pattern mining techniques has been commonly considered in data mining communitiesfor many years.The various efficient algorithms such as Apriori algorithms, FP-tree, SPADE, PrefixSpan,GST and SLPMiner [4], [5], [6], [7], and [8] have been proposed. Patterns can be discoveredby data mining techniques similar to sequential pattern mining , closed pattern mining [2] andfrequent item set mining,. To conquer the drawbacks of sequential patterns

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

and closed patterns,taxonomy models have been developed in pattern discovery technique [18].Feature classificationis assigning different task according to predefine group of documents. There is numerousclassification methods, such as Rocchio, Naive Bayes, KNN and SVM have been used in InformationRetrieval [14], [15], [16]. SVM is one of the main classification strategies used inmachine learning domain [14]. The grouping issues incorporate the single and multi-markedissue.Term based model documents having semantic meaning and documents are analyzed onthe basis of the term. The regular arrangement [13] to the numerous named issues is to breakdown it into a few classifiers, where a classifier allocates two predefined classifications. Thetwo classifications are positive or negative classification. Term based technique suffer from theissue of polysemy and synonymy [10]. Polysemy implies a word has numerous meaning and synonymy implies different words having the same meaning. IR gave numerous term-basedstrategies to this test [2], [3]. The similar research was also available in [2], [11] for developinga new techniques of post-processing of pattern mining, pattern summarization, which groupedpatterns into some clusters. Further patterns in the same clusters are into a master pattern thatconsists of a set of terms which are composed into a term-weight distribution. It is still a challengingproblem for pattern-based technique to deal with low frequency patterns (noise).In summary, the existing methods for finding relevance features are divided into threeapproaches. The first approach considers feature terms that come out in both positive samplesand negative samples that are Rocchio-based models [19] and SVM [14]. The second approachis based on probabilistic based models [15] in which terms show or do not show in positivedocuments and negative documents which defines their importance. The third approach considersonly positive patterns from the documents [11].

III. SOFTWARE REQUIREMENT SPECIFICATION

A. User Classes and Characteristics

To design products that satisfy their target users, a deeper understanding is needed of their user characteristics and product properties in development related to unexpected problems that the user's faces every now and then while developing a project. The study will lead to an interaction model that provides an overview of the interaction between user characters and the classes.It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms).

B. Nonfunctional Requirements

We continued to develop the RFD model and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method. This paper presents an innovative model for relevance feature discovery. With pout any design consideration it will be difficult to specify the performance criteria. But if developer did not define particular criteria the it is like to lose the performance of the system. The criteria which are defined to meet the performance are: Response Time, Workload, Scalability, and Platform.

IV. IMPLEMENTATION STATUS

In this paper, we proposes an innovative technique for finding and classifying low-level terms based on both their appearancesin the higher-level features (patterns) and their specificityin a training set. It also introduces a method to selectirrelevant documents (so-called offenders) that are closed tothe extracted features in the relevant documents in order toeffectively revise term weights.Compared with other methods,the advantages of the proposed model include:

1. It discover both negative and positive patterns in text documents as higher level features and deploys them over low-level features(terms).
2. It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

V. COMPARISON BETWEEN EXISTING SYSTEM AND PROPOSED SYSTEM

Item	Existing System	Proposed System
Algorithms	1. SP Mining 2. FClustring 3. WFeatures	1. HLF Mining 2. NRevision 3. Top K algorithm
Accuracy	Low	High
Complexity	Low	High
Explanation	In Existing system for relevance feature discovery SP mining algorithm was used. FClustring algorithm Categories terms from the relevant document into three categories and WFeature calculate the weight of term. Finally display the relevant document.	In our Proposed system, we used HLF Mining algorithm which is helpful to overcome the limitations of term based approaches. We also used SP mining algorithm for increase the accuracy of relevance feature discovery. After grouping the terms into three categories, the next step is to review the weight of the terms based on specific score. Ones we got a terms weight then we used different algorithms for ranking the document or terms. We used any one of the following algorithms for ranking. 1. Relevancy Ranking 2. TF-IDF 3. Top-k 4. Ranking SVM

VI. SYSTEM ARCHITECTURE

In this paper, we proposes an innovative technique for finding and classifying low-level terms based on both their appearancesin the higher-level features (patterns) and their specificityin a training set. It also introduces a method to selectirrelevant documents (so-called offenders) that are closed tothe extracted features in the relevant documents in order toeffectively revise term weights.Compared with other methods,the advantages of the proposed model include:

3. It discover both negative and positive patterns in text documents as higher level features and deploys them over low-level features(terms).
4. It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

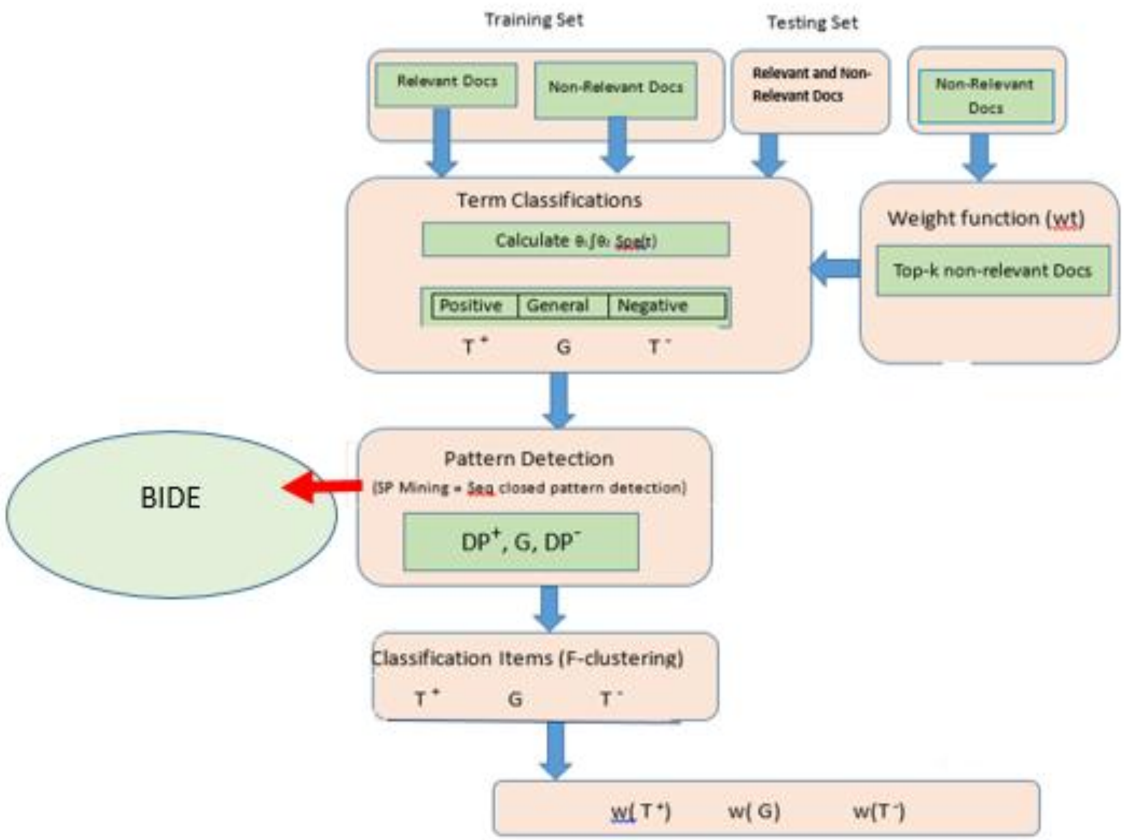

Fig.1 System Architecture.

Proposed system flow

1. Generate set of relevant and irrelevant documents from the RCV1 dataset.
2. Extract sequential pattern from relevant document.
5. Categories terms from the relevant document into three categories and rank the documentsaccording to the importance of category terms.

VII.ALGORITHM FOR RELEVANT FEATURE DISCOVERY USING TEXTMINING

The mainly the all system depend on efficient text document. Efficient Algorithms playimportant role in the relevant feature discovery from text document using text mining.The following steps explain the relevance feature of text documents:

1. Start.
2. Select the folder contain all documents.
3. User decides the term extraction with minimum value.
4. Perform term support weight calculation for all documents.
5. Document ranking using efficient algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

6. Assign term class specification using clustering algorithm.
7. Stop.

VIII. MATHEMATICAL MODEL

Let, S be the System having Input, Functions and Output. It can be represented as,
S = { I, F, O }
Where,
I is a set of all inputs given to the System,
O is a set of all outputs given by the System,
F is a set of all functions in the System.
• I =Text Documents
I1 = a1,a2,a3....n (relevant documents)
I2 = u1,u2,u3....n(non-relevant documents)
• I=I1 U I2
• F = F1, F2,F3,F4
• F1=spe(t) (Find specificity of term)
Where,
a= minimum boundary of general terms
b= maximum boundary of general terms
19
• F2=SPMining(D+; minsup)(Clustering of Patterns on the basis ofspecificity)
D+isallrelevantdocuments
• F3=FClustering(T,DP+,DP-,spe) (sequential pattern mining)
where,
T = a single cluster of all documents
DP+ = PositivePatterns
DP- = NegativePatterns
F4=WFeature(Updated training set, Extracted Features, term initial weight
function w)
• F={F1 U F2 U F3 U F4}
• O = {Positive Terms, General Terms, Negative Terms}

IX. SOFTWARE REQUIREMENT SPECIFICATION

In proposed work is designed to implement above software requirement. To implement this design following software requirements are used.

- Operating system: Windows XP/7.
- Coding Language :JAVA/J2EE
- Database : MYSQL
- Tool : Eclipse Luna

X. EXPERIMENTAL SET UP AND RESULT TABLE

1. Result Table

Comparison Results of RFD_1 and RFD_2 Models in all Assessing Topics on RCV_1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Model	Top-20	b/p	MAP	$F_{B=1}$	IAP
RFD_1	0.5570	0.4724	0.4932	0.4696	0.5125
RFD_2	0.5610	0.4729	0.4930	0.4696	0.5136
%chg	0.71%	0.11%	-0.04%	0.06%	0.21%

2. Result Evaluation

This paper also includes a set of experiments on RCV1 (TREC topics), Reuters-21578 and LCSH ontology. Theseexperiments illustrate that the proposed model achieves the best performance for comparing with term-based baselinemodels and pattern-based baseline models. The results also show that the term classification can be effectively approximated by the proposed feature clustering method, the proposed function is reasonable and the proposed modelsare robust. The experimental resultsdemonstrate that we can roughly choose the same amount of positive specific terms and general terms, and assign large weights to the positive specific terms.

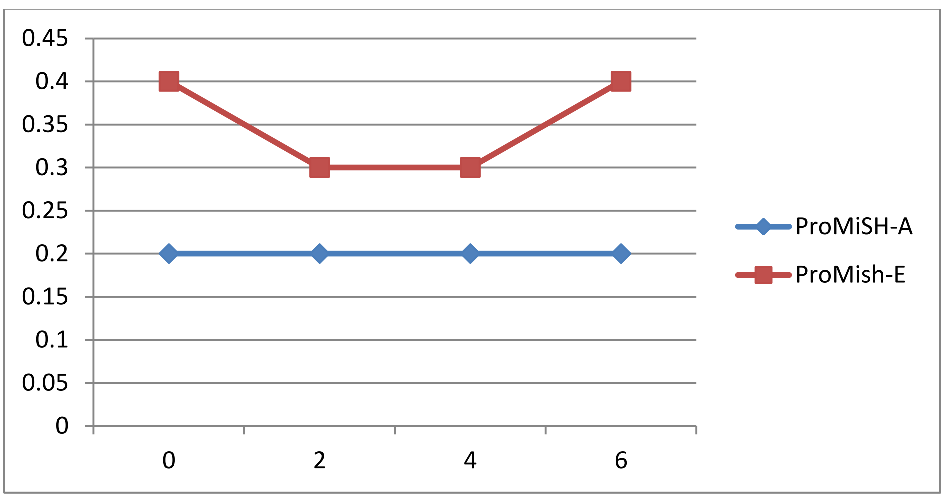

Fig. 02 Comparison for using different combinations of categories of terms for RFD2.

XI.CONCLUSION

The research proposes an alternative approach for relevance feature discovery in text documents. It presents a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity. It also introduces a method to select irrelevant documents for weighting features. In this paper, we continued to develop the RFD model and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method. The first RFD model uses two empirical parameters to set the boundary between the categories. It achieves theexpected performance, but it requires the manually testing of a large number of different values of parameters. The new model uses a feature clustering technique to automatically group terms into the three categories. Compared with the first model, the new model is much more efficient and achieved the satisfactory performance as well. This paper also includes a set of experiments on RCV1 (TREC topics), Reuters-21578 and LCSH ontology. These experiments illustrate that the proposed model achieves the best performance for comparing with term-based baseline models and pattern-based baseline models. The results also show that the term classification can

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

be effectively approximated by the proposed feature clustering method, the proposed spe function is reasonable and the proposed models are robust. This paper demonstrates that the proposed model was thoroughly tested and the results prove that the proposed model is statistically significant. The paper also proves that the use of irrelevance feedback is significant for improving the performance of relevance feature discovery models. It provides a promising methodology for developing effective text mining models for relevance feature discovery based on both positive and negative feedback.

REFERENCES

[1] Jaillet, S., Laurent, A., Teisseire, and M: Sequential patterns for text categorization. IntelligentData Analysis 10 (3), 199214 (2006).

[2] Wu, S., Li, Y., Xu, Y., P. Chen and B. Pham: Automatic pattern-taxonomy extraction for web mining. In: 3th IEEE/WIC/ACM WI International Conf. In Web Intelligence, pp.242248 (2004).

[3] Zhong, N., Li, Y., Wu, S.: Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, 24(1):30 44,,2011

[4] D.B. Liu. Web data mining: exploring hyperlinks, contents, and usage data. Data-centric systems and applications. Springer, Berlin, 2007.

[5] A. Rakesh and R. Srikant.Mining sequential patterns.In proceedings of the 11th InternationalConference on Data Engineering, pages 3.14, 1995.

[6] R. Afshar, X. Yan, and J. Han.Clospan: Mining closed sequential patterns in large data sets. In Data Mining (SDM03), pages 166.177, 2003.

[7] J. Han and K. Chang.Data mining for web intelligence. IEEE Computer, 35 (11): 64:70,2002.

[8] M. Zaki. Spade: an efficient algorithm for mining frequent sequences. In Machine Learning Journal, special issue on Unsupervised Learning, pages 31-60, 2001.

[9] Y. Xu and S. T. Wu, Y. Li, Deploying Approaches for Pattern Refinement in Text Mining, Proc. IEEE Sixth Intl Conf. Data Mining (ICDM 06), pp. 1157-1161, 2006.

[10] C. Buckley and G. Salton, Term-Weighting Approaches in Automatic Text Retrieval, InformationProcessing and Management: An Intl J., vol. 24, no. 5, pp. 513-523, 1988.

[11] N. Zhong and Y. Li, A. Agony. Mining positive and negative patterns for relevance feature discovery. In Proceedings of KDD10pages 753762, 2010.

[12] C. C. Yang. Search engine information retrieval in practice. J. Am. Soc. Inf. SCI. Technol., 61:430430, 2010.

[13] C. D. Manning, P. Raghavan, and H. Sects. Introduction to Information Retrieval.CambridgeUniversity Press, 2009.

[14] D. D. Lewis, Y. Yang, F. Li. Rcv1 and T. G. Rose: A new benchmark collection for text categorization research. J. Mach. Learn. Res., 5:361397, December 2004.

[15] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data.In Proceedingsof IJCAI03, pages 587592, 2003.

[16] X. L. Li, S. K. Ng and B. Liu. Learning to classify documents with only a small positive training set. In Proceedings of ECML07, pages 201213, Berlin, Heidelberg, 2007.

[17] S. E. Robertson and I. Soboroff. The trek 2002 filtering track report. In Proceedings of TREC02, 2002.

[18] Y. Li, X. Zhou, P. Bruce, R. Y. Lau and Y. Xu. Two-stage Decision Model for Information Filtering. Decision Support Systems, 52 (3): 706-716, 2012.

[19] T. Joachims. A probabilistic analysis of the rich algorithm with tfidf for text categorization.In Proc.On ICML97, pages 143151, 1997.

[20] N. Azam, and J. Yao. Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications, 39(5):47604768, 2012.

[21] R. Bekkerman and M. Gavish. High-precision phrase-based document classification on a modern scale. In Proc. of KDD11, pages 231239, 2011.

[22] G. Ifrim, G. Bakir, and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In Proceedings of KDD08:, pages 354362, 2008.

[23] X. Geng, T.-Y.Liu, T. Qin, A. Arnold, H. Li, and H.-Y.Shum. Query dependent ranking using k-nearest neighbor. In Proceedings of SIGIR08, pages 115122, 2008.

[24] Y. Gao, Y. Xu, and Y. Li. Topical Pattern Based Document Modelling and Relevance Ranking. In Proceedings of WISE(1)14, pages 186201, 2014.

[25] R. Sharma and S. Raman.Phrase-based text representation for managing the web documents.In International Conference on Information Technology: Coding and Computing, 165169, 2003.