Pattern Discovery For Text Mining Using Pattern Taxonomy

Miss Dipti S.Charjan[#1], Prof. Mukesh A.Pund [*2]

[1]M.E.(Scholar),Computer science & Engg., Department, PRMIT&R, Badnera , Sant Gadge Baba Amravati University,India.

[2]Associate Prof., Computer science & Engg., Department, PRMIT&R, Badnera , Sant Gadge Baba Amravati University,India

***Abstract*—In this paper, we focused on developing efficient mining algorithm for discovering patterns from large data collection. and search for useful and interesting patterns. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as frequent itemsets, closed frequent itemsets, co-occurring terms. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. In proposed system we can take sufficient .txt file as inputs & we apply various algorithms & generate expected results.**

Text-mining refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. An important difference with search is that search requires a user to know what he or she is looking for while text mining attempts to discover information in a pattern that is not known beforehand.

***Keywords*—Text mining, text classification, pattern mining, pattern evolving, information filtering**

I. Introduction

Text mining is the discovery of interesting knowledge in text documents. It is challenging issue to find accurate knowledge in text documents to help users to find what they want. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be effectively use and update discovered patterns and apply it to field of text mining [17][31]. Data mining is therefore an essential step in the process of knowledge discovery in databases, which means data mining is having all methods of knowledge discovery process and presenting modeling phase that is application of methods and algorithm for calculation of search pattern or models. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame [2]. With a large number of patterns generated by using the data mining approaches, how to effectively exploit these patterns is still an open research issue.

Text mining is the technique that helps users find useful information from a large amount of digital text data [16]. It is therefore crucial that a good text mining model should retrieve the information that users require with relevant efficiency. Traditional Information Retrieval (IR) has the same objective of automatically retrieving as many relevant documents as possible whilst filtering out irrelevant documents at the same time. However, IR-based systems do not adequately provide users with what they really need. Many text mining methods have been developed in order to achieve the goal of retrieving for information for users. We 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. The process of knowledge discovery may consist as following:

- Data Selection
- Data Processing
- Data Transaction
- Pattern Discovery
- Pattern Evaluation.

Text mining is also called as knowledge discovery in databases because, we frequently find in literature text mining as a process with series of partial steps among other things also information extraction as well as the use of data mining. when we analyze data in knowledge discovery in databases is aims of finding hidden patterns as well as connections in those data. While the ability to search for keywords or phrases in a collection is now widespread such search only marginally supports discovery because the user has to decide on the words to look for. On the other hand, text mining results can suggest “interesting” patterns to look at, and the user can then accept or reject these patterns as interesting. In this research we present pattern taxonomy model which extracting descriptive frequent patterns by pruning the meaningless ones. patterns are sorted based on their repeations.

II.Literature Review

A.Text Mining

Text mining is nothing but data mining, as the application of algorithm as well as methods from the machine learning and statistics to text with goal of finding useful pattern, Whereas data mining belongs in the corporate world because that's where most databases are, text mining promises to move machine learning technology out of the companies and into the home" as an increasingly necessary Internet adjunct (Witten & Frank, 2000) – i.e., as "web data mining" (Hearst, 1997). Laender, Ribeiro-Neto, da Silva, and Teixeira (2001) provide a current review of web data extraction tools.

Text mining is also referred to as text data mining, roughly equivalent to text analytics, it refers to process of deriving high quality information form text. and high quality of information is derived through devising of patterns. Text analysis involves information retrieval, lexical analysis, word frequency distributions, pattern recognition, information extraction, and data mining techniques including link and association analysis, visualization to turn text into data for analysis via..natural language processing and analytical methods. On otherhand we called -Text mining is a variation on field called data mining, that tries to find interesting patterns from large datasets. This is a concept of text mining describe in this section.

B. Pattern Discovery

The pattern used as a word or phase that is extracted from the text document. There are numbers of patterns which may be discovered from a text document, but not all of them are interesting. Only those evaluated to be interesting in some manner are viewed as useful knowledge. It is midfield task between association rule mining and inductive learing. It aims at finding patterns in labelled data that are descriptive.

A system may encounter a problem where a discovered pattern is not interesting a user. Such patterns are not qualified as knowledge. Therefore, a knowledge discovery system should have the capability of deciding whether a pattern is interesting enough to form knowledge in the current context.

C. Pattern Taxonomy

Pattern can be structured into taxonomy-used knowledge discovery model is developed towards applying data mining techniques to practical text mining applications. Knowledge Discovery in Databases (KDD) can be referred to as the term of data mining which aims for discovering interesting patterns or trends from a database. In particular, a process of turning low-level data into high-level knowledge is denoted as KDD. The concept of KDD process is the data mining for extracting patterns from data. we focus on development of knowledge discovery model to effectively use & update discovered patterns and apply it to the field of text mining.

III. Praposed System

In terms of pattern discovery, the data mining techniques can be used for pattern discovery.

In Fig.1 we pass input file type .txt .& read that text file.

Then we apply various algorithms on it like stemmer algorithem, PTM and IPE & display result. However, the main drawback of using data mining is the explosion of numbers of discovered patterns. Both closed pattern-based approaches and non-closed-based approaches can be adopted and used in a pattern-based.

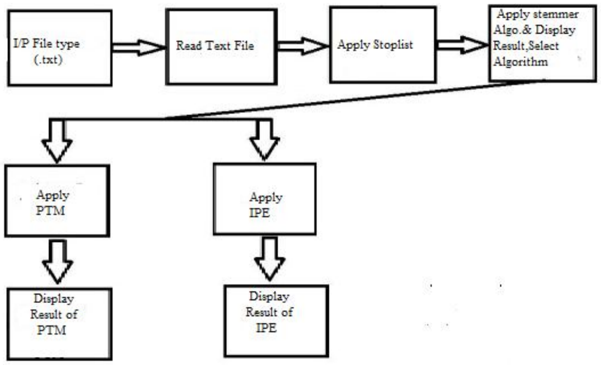


Fig. 1 Proposed System Block Diagram.

IF system for pattern discovery. The weight of a pattern is in direct proportion to the pattern's frequency in documents.

A. Pattern Taxonomy Model

There are two main stages are consider in PTM ,first one is –how to extract useful phases from text documents. and second one is, how to use these discovered patterns to improve effectiveness of a knowledge discovery system. The main focus of this algorithm is deploying process, which consist of pattern discovery and term support evaluation. In this paper, we assume that all text documents are split into paragraphs. So a given document d yields a set of paragraphs PS(d). Let D be a training set of documents, which consists of a set of positive & negative documents, Let T = {$t_{1.}$ $t_{2.}$ $t_{3.}$ $t_{4....}$ $t_{m.,\,..}$} be a set of terms (or keywords) which can be extracted from the set of positive documents [8][31].

B. Frequent and Closed Patterns

Frequent Patterns is one that occurs in atleast a user specific percentage of database, that percent is called support

Given a termset X in document d, X is used to denote the covering set of X for d, which includes all paragraphs dp $\in$ PS(d) such that X $\subseteq$ dp. i.e.

X ={ dp | dp $\in$ PS (d) , X $\subseteq$ dp }

Its absolute support is the number of occurrences of X in PS(d), that is $\sup_\alpha$ (X) = |X|. Its relative support is thefraction of the paragraphs that contain the pattern, that is, $\sup_r$ (X) = $\frac{X}{PS(d)}$.

A termset X is called frequent pattern if its $\sup_r$ (or supα) $\geq$ min_sup.[31]

C. Closed Sequential Patterns

Closed sequential pattern is a frequent sequential pattern such that it is not included in another sequential pattern having exactly same support.

A sequential pattern s=<t1; . . . ; tr>(ti elements of T) is an ordered list of terms. A sequence s1= <x1; . . . ; xi> is a subsequence of another sequence s2<=y1; . . . ; yj>, is called s1 is sub-set of s2, iff j_1; . . . ; j_y such that 1<= j_1 < j_2 . . . < j_y <=j and x_1=y_{j1}; x_2=y_{j2}; . . . ; x_i=y_{jy}. Given s1 is sub-set of s2; we usually say s1 is a sub-pattern of s2, and s2 is a super pattern of s1. In the following, we simply say patterns for sequential patterns.[31]

A sequential pattern X is called frequent pattern if its relative support (or absolute support) _ min sup, a minimum support.[2].A frequent sequential pattern X is called closed if not any super pattern X1 of X such that $\sup_a$(X1)= $\sup_a$(X).

- Composition Operation

Let p1 and p2 be sets of term number pairs. P1$\oplus$P2 is called composition of p1 and p2 which satisfies-

P1$\oplus$P2 = {(t,x1+x2)|(t,x1)€P1,(t,x2) €P2} U {(t, x)|(t, x) € P1UP2, not(t,_) €P1$\cap$P2}

Where is the wild card that matches any number

Example:

{(t1, 3),(t2, 2), (t3, 3), (t4, 3)}$\oplus$ {(t2,3), (t5,4)} = {(t1, 3),(t2, 5), (t3, 3), (t4, 3), (t5, 4)}.

Here we add the common elements and which is not common we write as it is. In the above example t2 elements are available in both sets so {(t2,2+3)} as composition and another elements of sets we write as it is.[31]

D. Inner Pattern Evolution

In this section, we discuss how to reshuffle supports of terms within normal forms of d-patterns. The technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evolution here, because it only changes a pattern's term supports within the pattern. A threshold is usually used to classify documents into relevant or irrelevant categories. Using the d-patterns, the threshold can be defined naturally as follows:

$\text{Threshold}(dp) = \min_{p \, E \, DP} (\sum_{tw \, E \, \beta \, p} \text{support}(t))$[31]

E. Shuffling

The time complexity of Algorithm decided by the number of calls for Shuffling algorithm and the number of using ⊕ operation.

The task of algorithm Shuffling is to tune the support.

Distribution of terms within a d-pattern. A different strategy is dedicated in this algorithm for each type of offender. As stated in steps in the algorithm Shuffling, complete conflict offenders (d-patterns) are removed since all elements within the d-patterns are held by the negative documents indicating that they can be discarded for preventing interference from these possible "noises."

IV. Experimental Analysis

A. Requirement Analysis

For implementation of this system, we used .Net technology. A main part of the .Net technology and structure is the ASP.net set of technologies. These web development technologies are used in the making of Websites and net services working on the .NET infrastructure. ASP.NET was billed by Microsoft from one of their big technologies and web programmers can make use of any encoding language they want to write ASP.NET, from Perl to C Sharp (C#) and of course VB.NET and a few extra language unspoken with the .NET technology.

B. Hardware And Software Requirements

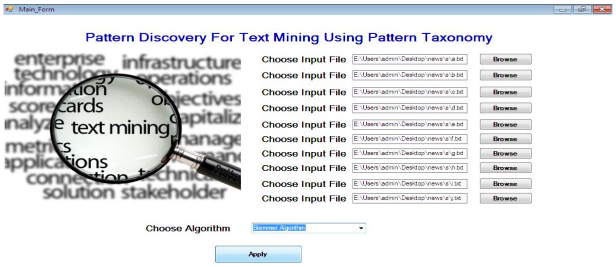
1) Hardware Requirements

- Windows XP
- RAM – 1GB
- Hard Disk - 40GB

2) Software Requirements

- .Net Framework
- IIS 7.0
- SQL Server

C. Result



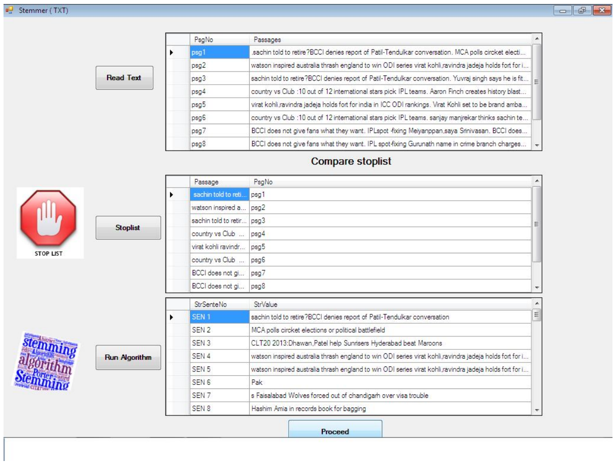
Main_Form

Pattern Discovery For Text Mining Using Pattern Taxonomy

Choose Input File	E:\Users\admin\Desktop\news\s\a.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\b.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\c.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\d.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\e.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\f.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\g.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\h.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\i.txt	Browse
Choose Input File	E:\Users\admin\Desktop\news\s\j.txt	Browse

Choose Algorithm: Stemmer Algorithm

Fig.2 selection of inputs & apply Stemmer Algorithm



Stemmer (TXT)

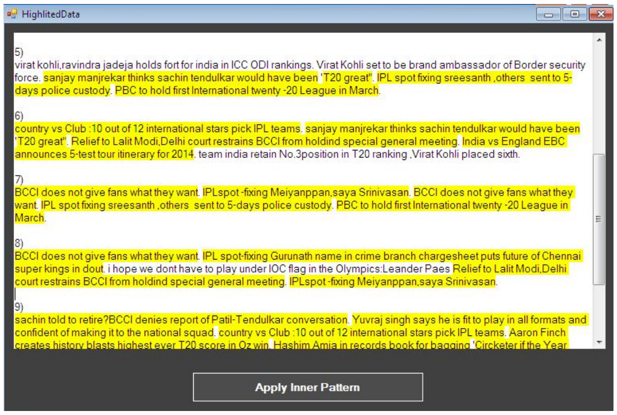
PsgNo	Passages
psg1	.sachin told to retire?BCCI denies report of Patil-Tendulkar conversation. MCA polls circket electi...
psg2	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
psg3	sachin told to retire?BCCI denies report of Patil-Tendulkar conversation. Yuvraj singh says he is fit...
psg4	country vs Club :10 out of 12 international stars pick IPL teams. Aaron Finch creates history blast...
psg5	virat kohli,ravindra jadeja holds fort for india in ICC ODI rankings. Virat Kohli set to be brand amba...
psg6	country vs Club :10 out of 12 international stars pick IPL teams. sanjay manjrekar thinks sachin te...
psg7	BCCI does not give fans what they want. IPLspot -fixing Meiyanppan,saya Srinivasan. BCCI does...
psg8	BCCI does not give fans what they want. IPL spot-fixing Gurunath name in crime branch charges...

Compare stoplist

Passage	PsgNo
sachin told to reti...	psg1
watson inspired a...	psg2
sachin told to retir...	psg3
country vs Club ...	psg4
virat kohli ravindr...	psg5
country vs Club ...	psg6
BCCI does not gi...	psg7
BCCI does not gi...	psg8

StrSenteNo	StrValue
SEN 1	sachin told to retire?BCCI denies report of Patil-Tendulkar conversation
SEN 2	MCA polls circket elections or political battlefield
SEN 3	CLT20 2013:Dhawan,Patel help Sunrisers Hyderabad beat Maroons
SEN 4	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
SEN 5	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
SEN 6	Pak
SEN 7	s Faisalabad Wolves forced out of chandigarh over visa trouble
SEN 8	Hashim Amia in records book for bagging

Fig.3 Stemmer Algorithm.



HighlitedData

5)
virat kohli,ravindra jadeja holds fort for india in ICC ODI rankings. Virat Kohli set to be brand ambassador of Border security force. sanjay manjrekar thinks sachin tendulkar would have been 'T20 great". IPL spot fixing sreesanth ,others sent to 5-days police custody. PBC to hold first International twenty -20 League in March.

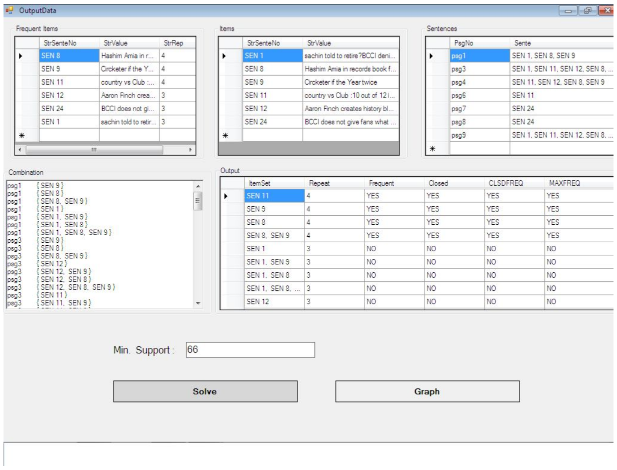
6)
country vs Club :10 out of 12 international stars pick IPL teams. sanjay manjrekar thinks sachin tendulkar would have been 'T20 great". Relief to Lalit Modi,Delhi court restrains BCCI from holdind special general meeting. India vs England EBC announces 5-test tour itinerary for 2014. team india retain No.3position in T20 ranking ,Virat Kohli placed sixth.

7)
BCCI does not give fans what they want. IPLspot -fixing Meiyanppan,saya Srinivasan. BCCI does not give fans what they want. IPL spot fixing sreesanth ,others sent to 5-days police custody. PBC to hold first International twenty -20 League in March.

8)
BCCI does not give fans what they want. IPL spot-fixing Gurunath name in crime branch chargesheet puts future of Chennai super kings in dout. i hope we dont have to play under IOC flag in the Olympics:Leander Paes Relief to Lalit Modi,Delhi court restrains BCCI from holdind special general meeting. IPLspot -fixing Meiyanppan,saya Srinivasan.

9)
sachin told to retire?BCCI denies report of Patil-Tendulkar conversation. Yuvraj singh says he is fit to play in all formats and confident of making it to the national squad. country vs Club :10 out of 12 international stars pick IPL teams. Aaron Finch creates history blasts highest ever T20 score in Oz win. Hashim Amia in records book for bagging 'Circketer if the Year

Fig.4 PTM / Inner Pattern Algorithm



Frequent Items

	StrSenteNo	StrValue	StrRep
▸	SEN 8	Hashim Amia in r...	4
	SEN 9	Circketer if the Y...	4
	SEN 11	country vs Club :...	4
	SEN 12	Aaron Finch crea...	3
	SEN 24	BCCI does not gi...	3
	SEN 1	sachin told to retir...	3
*			

Items

	StrSenteNo	StrValue
▸	SEN 1	sachin told to retire?BCCI deni...
	SEN 8	Hashim Amia in records book f...
	SEN 9	Circketer if the Year twice
	SEN 11	country vs Club :10 out of 12 i...
	SEN 12	Aaron Finch creates history bl...
	SEN 24	BCCI does not give fans what ...
*		

Sentences

	PsgNo	Sente
▸	psg1	SEN 1, SEN 8, SEN 9
	psg3	SEN 1, SEN 11, SEN 12, SEN 8, ...
	psg4	SEN 11, SEN 12, SEN 8, SEN 9
	psg6	SEN 11
	psg7	SEN 24
	psg8	SEN 24
	psg9	SEN 1, SEN 11, SEN 12, SEN 8, ...
*		

Combination

```
psg1    { SEN 9 }
psg1    { SEN 8 }
psg1    { SEN 8,  SEN 9 }
psg1    { SEN 1 }
psg1    { SEN 1,  SEN 9 }
psg1    { SEN 1,  SEN 8 }
psg1    { SEN 1,  SEN 8,  SEN 9 }
psg3    { SEN 9 }
psg3    { SEN 8 }
psg3    { SEN 8,  SEN 9 }
psg3    { SEN 12 }
psg3    { SEN 12,  SEN 9 }
psg3    { SEN 12,  SEN 8 }
psg3    { SEN 12,  SEN 8,  SEN 9 }
psg3    { SEN 11 }
psg3    { SEN 11,  SEN 9 }
```

Output

	ItemSet	Repeat	Frequent	Closed	CLSDFREQ	MAXFREQ
▸	SEN 11	4	YES	YES	YES	YES
	SEN 9	4	YES	YES	YES	YES
	SEN 8	4	YES	YES	YES	YES
	SEN 8, SEN 9	4	YES	YES	YES	YES
	SEN 1	3	NO	NO	NO	NO
	SEN 1, SEN 9	3	NO	NO	NO	NO
	SEN 1, SEN 8	3	NO	NO	NO	NO
	SEN 1, SEN 8, ...	3	NO	NO	NO	NO
	SEN 12	3	NO	NO	NO	NO

Min. Support : 66

Fig.5 Frequent and Close Patterns.

D. System Performance

Fig.6 shows the size of the various Pattern.. The Y-axis represents the number of times that pattern may occurs. The X-axis represent the number of patterns in the form of sentences.

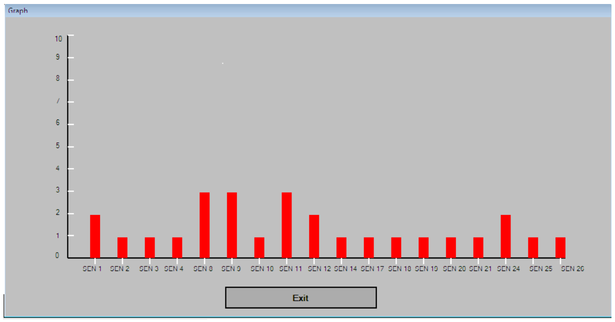


Fig.6 System Performance of Each Pattern

V.CONCLUSION AND FUTURE WORK

Many data mining techniques have been proposed in the last decade, these techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective, because some useful long patterns with high specificity lack in support (i.e., the low-frequency problem).In this research work, have mainly focused on developing efficient mining algorithm for discovering patterns from a large data collection. and search for useful and interesting patterns. In proposed technique we can take input file .txt then we apply various algorithms such as stemmer, PTM, Inner pattern & display expected output. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

In our future work we will investigate better means of exploration of long patterns and look at more diverse kinds of texts, especially large collections of text where a two level hierarchy may not be sufficient. We will also support the filtering of patterns by their usage trend over time. Metrics can be defined to characterize frequency distributions associated with each pattern and identify that are increasing, decreasing, showing spikes or gaps, etc. Finally, we have focused here on patterns of repetitions, other features can be extracted from the text (e.g. name entities, part of speech patterns) and explored in a similar fashion.

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

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

[30] I. 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] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining",vol.24,No.1,Jan.2012.