International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-2 Issue-3, July 2013

Spam Detection Using Baysian with Pattren Discovery

Asmeeta Mali

Abstract- ***Text mining is nothing but the discovery of interesting knowledge in text documents. But there is a big challenging issue that how to guarantee the quality of discovered relevant features. And that are in the text documents for describing user preferences because of the large number of terms, patterns and noise. For text mining there are basically two types of approaches; one is term based approach and another is phrase based approach. But term based approach suffered with the problem of polysemy and synonymy. And phrase based approach suffered with low frequency occurrence. But phrase based approachs are better than the term based approachs. But pattern based approach is better than the term based and phrase based approach. The proposed method is an innovative and effective pattern discovery technique. This method includes two main processes pattern deploying and inner pattern evaluation. This paper presents an effective technique to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Using Baysian filtering algorithm and effective pattern Discovery technique we can detect the spam mails from the email dataset with good correctness of term.***

Index Terms— Text mining, information filtering, pattern mining, sequential pattern, closed sequential patterns.

I. INTRODUCTION

Now days, due to rapid growth of digital data it is difficult to find useful information and knowledge. So we have to draw attention towards knowledge discovery and data mining. Turning such data into the useful information and knowledge we have to pay great deal of attention with coming need. There are so many applications like market analysis and business management can have benefit by the knowledge discovery and data mining. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large database. Data mining is an essential step in the process of knowledge discovery in database.In past decade various data mining techniques have been presented. These techniques included association rule mining, sequential pattern mining frequent itemset 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.

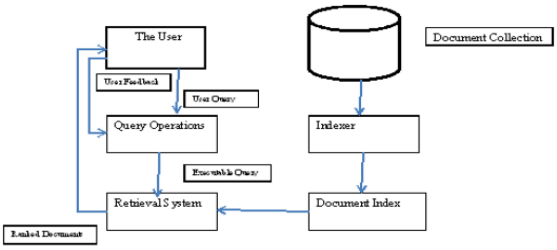
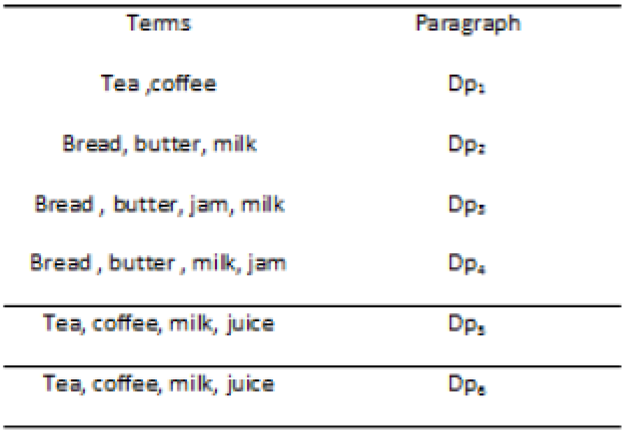


Figure 1: A General IR System architecture

Data mining approaches generated a large number of patterns. But still there is an open research issue that how to effectively use and update these patterns. In this paper, we focus on the development of knowledge. Discovery model is used to effectively use and update the discovered patterns and apply it to the field of text mining. Text mining is the discovery of interesting knowledge and information in text documents. It is a challenging issue to find accurate knowledge and information in text documents to help users to find what they want.

Before text mining, information retrieval provided many term based approaches to solve this challenge. But there is one advantage for term based methods in text mining that they give efficient computational performance as well as mature theories for term weighting. But there are some problems with this term based methods in text mining that are polysemy and synonymy. polysemy means a word has multiple meanings and synonymy means a multiple words having the same meaning.

In the field of text mining preprocessing of data plays an important role. If we use normal text documents as input to text mining methods then following tasks should perform on those documents. Stop word removal, stemming, handling of digits, hyphens, punctuations, cases of letters. If we use the we documents then there are addition of two tasks in previous task set, HTML tag removal and identification of main content block. Stop word Removal Stop words are the words which are frequently occurring and insignificant words helps to constructs the sentences in the language. Example like a, and the what, there, that, also, etc. Stemming Removal suffix and stripping. Stem is the portion of word which left after removing the suffix and stripping of that word. Example Computing, Computer, Computation are reduced into Compute.



Terms	Paragraph
Tea ,coffee	Dp_1
Bread, butter, milk	Dp_2
Bread , butter, jam, milk	Dp_3
Bread , butter , milk, jam	Dp_4
Tea, coffee, milk, juice	Dp_5
Tea, coffee, milk, juice	Dp_6

Figure 2: Table 1 :A set of Paragraphs

Revised Manuscript Received on 30 July 2013.
* Correspondence Author
AsmeetaMali. DYPIET, University of Pune, Maharashtra, India.

II. RELATED WORK

In the past, many text representations have been proposed. A well-known representation is the bag of words that uses keywords or terms as elements in the vector of the features space. In Rpcchio classifiers, the tf*idf weighting scheme is used for text representation. For many years, Pattern mining has been extensively studied in data mining communities. There are varieties of efficient algorithms such as Apriori I, Prefix span, FP-tree, SPADE, SLPMINER and GST have been proposed for pattern mining. In the field of text mining there was still an open problem that searching for useful and interesting patterns and rules. Pattern mining techniques can be used find various text patterns such as sequential patterns, frequent item sets, co-occurring terms and multiple grams for building up e presentation with this new type features. But the challenging issue is the how to effectively deal with large amount of discovered patterns for the challenging issue closed sequential pattern have been

Used for text mining in it is proposed that the concept of closed pattern in the text mining was useful and had the potential for improving the performance of text mining. Pattern taxonomy model was developed to improve the effectiveness by effectively using closed patterns in text mining.

There are mainly two types of approaches used for text Mining, term based and another is phrase based. But Both approaches have some problems. Rocchio ,BM25, SVM are the example of term based approach. These methods are widely used for information retrieval. BM25 method is used in the Okapi Model. Okapi model is based on the aforementioned probabilistic models. This method uses the term frequency and the document length to calculate the support.

Rocchio Method is the one of the early and effective Algorithm for information retrieval. This method is based on the relevance feedback approach. It uses the User identified relevant and irrelevant documents to expand the original query. New query is then used to perform retrieval again. In this method relevant documents are considered more important than the irrelevant documents. This method is simple and efficient to compute, and usually produce good results.

SVM methods are basically a machine learning methods. There are two types of learning methods labeled learning and positive learning. SVM method is positive and unlabeled learning method. It gives the better results than the Rocchio and BM25 methods. But it also has some drawback to overcome that use pattern based methods for information retrieval. There are so many pattern based methods used for the text mining like SPM means Sequence pattern mining and PTM which is explained in detail in next section.

PTM: In this paper we assume that all documents are Split into paragraphs given document having a set of Paragraphs let D be the training set of documents which contain the set of positive documents D+ and set of negative documents. T be the set of terms t which can be extracted from the set of positive documents.

Frequent and closed patterns:X is used to denote the convening set of X for d. **Absolute support** means the number of occurrences of X in PS(d)

Relative support means the fraction of the paragraphs that contain the pattern

Frequent pattern: The term set X is called frequent pattern if its relative or absolute support is greater than or equal to minimum support.

Table 1 shows documents and terms sets Table 2 shows 10 frequent patterns and there covering sets. But from table 2 not all frequent patterns are useful. For example pattern (bread, butter) always occur with the term milk in paragraphs. So(bread, butter) is a shorter pattern is always a part of larger pattern(bread, butter, milk).A pattern is closed if none of its immediate superset has the same support as the pattern .

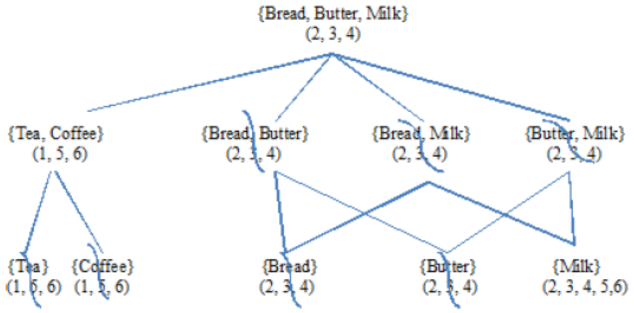
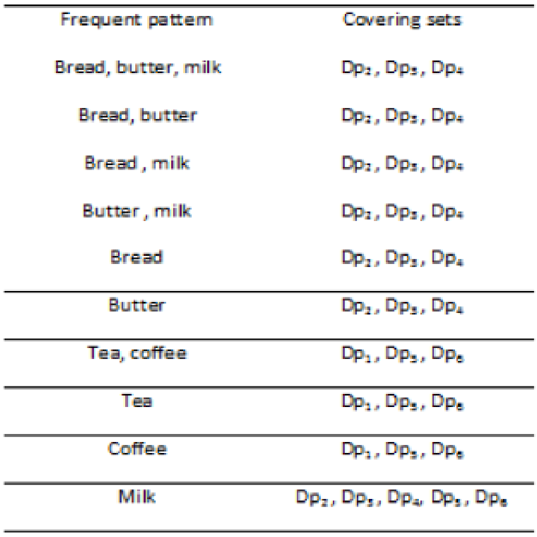


Figure 3: Pattern Taxonomy

Pattern taxonomy: Patterns can be structured into taxonomy by using a is a (or subset) relation. Tables 1 have set of paragraphs of documents. Table 2 have discovered ten frequent pattern assuming *minsup* =0.2. There are only three closed pattern in this example

(Bread, butter, milk), (tea, coffee), (milk).

Closed sequential pattern: The property of closed patterns can be used to define closed sequential patterns.



Frequent pattern	Covering sets
Bread, butter, milk	Dp_2 , Dp_3 , Dp_4
Bread, butter	Dp_2 , Dp_3 , Dp_4
Bread , milk	Dp_2 , Dp_3 , Dp_4
Butter , milk	Dp_2 , Dp_3 , Dp_4
Bread	Dp_2 , Dp_3 , Dp_4
Butter	Dp_2 , Dp_3 , Dp_4
Tea, coffee	Dp_1 , Dp_5 , Dp_6
Tea	Dp_1 , Dp_5 , Dp_6
Coffee	Dp_1 , Dp_5 , Dp_6
Milk	Dp_2 , Dp_3 , Dp_4, Dp_5 , Dp_6

Figure 4: Table 2 : Frequent patterns and covering sets

III. PROGRAMMER'S DESIGN

In the proposed method two important processes presented. Pattern deploying and inner pattern evaluation. In the pattern deploying method main focus is on the discovery of the patterns and the term support evaluation for calculating the patterns the composition operation is used and the support of the term.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In the inner pattern evaluation discuss how to reshuffle support of terms within normal forms of the patterns on the negative documents in the training sets. These techniques will be useful to reduce the side effects if the noisy patterns because of the low frequency problem.

3.1. Mathematical Model

In the proposed method we have used some mathematical formula for calculating support and threshold. And one composition operation for calculating d patterns. All these are given below:

1. To find out the patterns we use the composition operation. Let p1 and p2 be sets of term number pairs it denotes a single term and its support in document D. p1= (tea, 1), (coffee, 2), (bread, 3) and p2 = (coffee, 4) .

   Then p1 $\oplus$ p2 is given below:

   (tea, 1), (coffee, 2), (bread, 3) $\oplus$ (coffee, 4)
   = (tea, 1), (coffee, 6), (bread, 3)
2. To calculate the support of the term we use the Following formula:

$$Weight(D) = \sum_{t \in T} Support(t) \tau(t, d)$$

where support(t) is given by the d pattern mining algorithm and threshold τ (*t*, *d*) = 1 if (*t* $\in$ *d*) otherwise τ (*t*, *d*) = 0.

3. A threshold is usually used to classify documents into relevant or irrelevant categories.

$$Threshold\ (DP) = \min_{p \in DP} \left(\sum_{(t,\omega) \in \beta(p)} \text{support}\ (t) \right)$$

Using threshold value we can classify the documents into positive and negative sets. Using threshold we can find the off Enders. And categories them into two types 1) a complete conflict offenders 2) a partial conflict offenders. The basic idea behind the updating patterns is as follows: complete conflict offenders are removed from the d patterns first. For partial conflict offenders their term supports are reshuffled to reduce the effects of noise documents.

3.2. Dynamic Programming and Serialization

Dynamic programming and serialization: Dynamic programming is every powerful algorithmic paradigm in which problem is solved by identifying the collection of sub problems and taking them one by one smallest first using the answers to small problems to help figure out larger ones until the whole lot of them solved. In the proposed method uses the dynamic programming approach serialization establishes a logical ordering on all operations the resulting parallel execution is predictable and deterministic.

3.3. The Algorithm

The proposed method includes three algorithms. These are DP mining algorithm, IPEvolving algorithm and Shuffling algorithm.

DP mining algorithm DP means the discovered pattern mining algorithm. It describes the training process of finding the set of d-patterns which is also known as the discovered pattern. To improve the efficiency of the pattern taxonomy mining or DP mining, SPMining algorithm was proposed[11]to find all closed sequential patterns, which uses the apriori property.

Step 1 -Initialize set of d-patterns to zero.

Step 2-5 -Find out the closed sequential patterns using SPMining algorithm

Step 6-9 -All discovered patterns in the positive documents are composed into a d-pattern giving rise to a set of d-patterns.

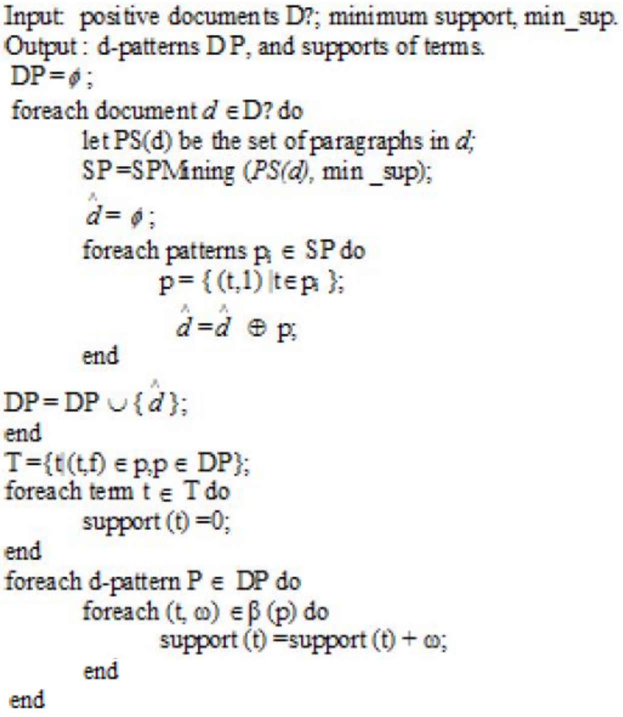
Step 12 – 19 - Term supports are calculated based on the normal form

for all terms in d-pattern. Detail algorithm is given in fig. 5.

IPEvolving algorithm It is also known as the Inner Pattern evolving algorithm. This technique is useful to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evaluation. 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. There are two types of offenders: Complete conflict offenders and Partial conflict offenders.

Step 1 -Initialize set of term support pairs to zero.

Step 2 -Estimate the threshold for finding the noise negative documents.



```
Input: positive documents D?; minimum support, min_sup.
Output: d-patterns DP, and supports of terms.
DP = $\phi$;
foreach document $d \in$ D? do
        let PS(d) be the set of paragraphs in $d$;
        SP =SPMining ($PS(d)$, min _sup);
        $\hat{d} = \phi$;
        foreach patterns $p_i \in$ SP do
                p = { (t,1) | t $\in p_i$ };
                $\hat{d} = \hat{d} \oplus p$;
        end
DP = DP $\cup$ { $\hat{d}$ };
end
T={t|(t,f) $\in$ p,p $\in$ DP};
foreach term t $\in$ T do
        support (t) =0;
end
foreach d-pattern P $\in$ DP do
        foreach (t, $\omega$) $\in \beta$ (p) do
                support (t) =support (t) + $\omega$;
        end
end
```

Figure 6: IPEvolving Algorithm

Step 5 - Calculate the base which is certainly not zero.

Step 6-9 - agian calculate the support for the terms.

The proposed work is divided into following modules :

The proposed model is divided into two phases 1. Training phase 2. Testing phase. In training phase the proposed model first calls the pattern mining algorithm to find the patterns in positive documents based on a *minsup* and evaluates term support by deploying the patterns to terms it also calls inner pattern evaluation algorithm to revise term n supports using noise negative documents based on experimental co efficient at. In testing phase it evaluate weights for all incoming documents using equation 2. The incoming documents then can be sorted based on these weights. When we have to detect the spam mails from the dataset first we have to apply the effective pattern discovery for that mail .

Then we get the effective patterns now we use these patterns for the baysian classification. Using baysian Classifier we calculate the probability of each effective pattern for training set of spam and ham mails. Based on that we decide that mail is spam mail or nonspam mail.

3.4. Data independence and Data Flow architecture

Data independence is the type of data transparency for Centralized DBMS. It references to the immunity of user applications to make changes in the definition and organization of data. There are two types of data independence.

1. Physical independence; it deals with hiding details of the storage structure from user application.
2. Logical independence: it is the logical structure of data known as schema definition.

Data flow architecture: Is a series of function in a computer system where each step is automatically generated by the actions of previous function. It is also known as reactive programming. Data flow architecture is considered to be fairly simple form of programming it often used by less experienced programmers.

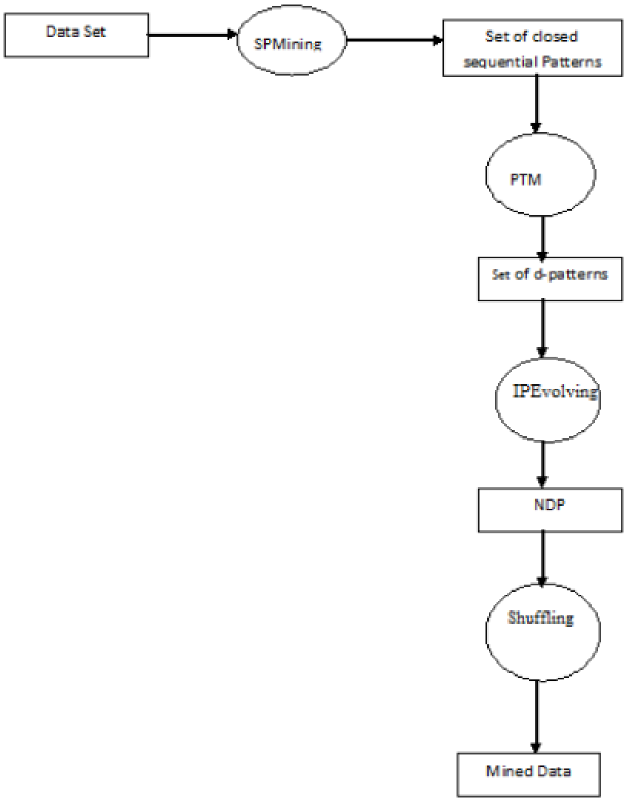


Figure 7: Process Flow Diagram

3.5. Baysian Classification

In the baysian classification we have to find out the probability of term in the spam mails or in the nonspam mails then we decide the spam probability for that term. In this proposed method we use the terms which are the output of the effective pattern discovery. There is no need to find the terms from that mail. Find out the probability for that term in spam mail and in nonspam mails also. Depend upon those two probabilities we can find out the spam probability for that mail. We can decide the input mail is spam or nonspam is depending upon the spam probability.

Select Dataset \project\text mining\dataset\e-mails\

Figure 8: Effective Pattern Discovery

Select Dataset \project\text mining\dataset\e-mails\

Figure 9: Spam Detection

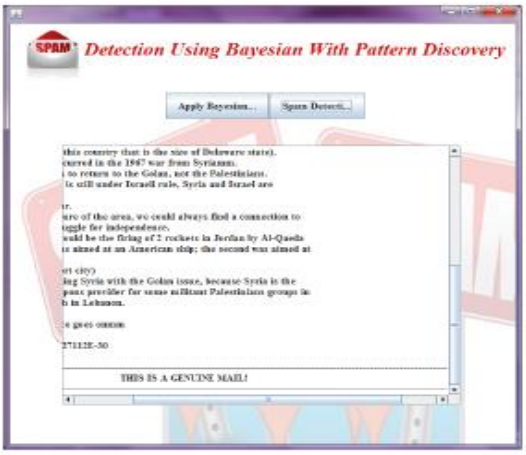
IV. RESULT AND DISSCUSSION

The proposed algorithm will be implemented using JAVA. In this we first implement effective pattern discovery technique then we implement the baysian classifier for classification of mails. In this proposed method we use both pattern discovery and baysian classification. We will compare the proposed algorithm with various term based, phrase based algorithm and state of the art algorithm. For the proposed method use the documents which are the text documents. We take that text document as an input. Then preprocess that document which contain the terms stemming and stopword removal technique can be used. For this we use porter algorithm is selected for suffix stripping several standard measures based on the precision and recall and are used the precision is the fraction of the retrieve documents that are relevant to the topic and recall is the fraction of relevant document that have been retrieved top k, b/p, fb measure, IAP, MAP, these are the other measures based on precision and recall. We get the effective pattern from that document using proposed technique.

Then we use those effective patterns for the baysian classifier for classification of that mail. So, using proposed approach we can get a better result than any baysian classifier.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The proposed approach is compared with other approaches like data mining based methods like sequential pattern mining closed sequential mining etc. Second category includes the concept based models like CBM pattern matching model and last category includes Rocchio, Probabilistic, Ngram and state of the art models, SVM, BM25. After comparing our results with above methods, proposed method is improved and effective output.



this country that is the size of Delaware state).
curred in the 1967 war from Syrians.
to return to the Golan, not the Palestinians.
is still under Israeli rule, Syria and Israel are

r.
are of the area, we could always find a connection to
uggle for independence.
ould be the firing of 2 rockets in Jordan by Al-Qaeda
s aimed at an American ship; the second was aimed at

rt city)
ing Syria with the Golan issue, because Syria is the
port provider for some militant Palestinian groups in
h in Lebanon.

e goes on

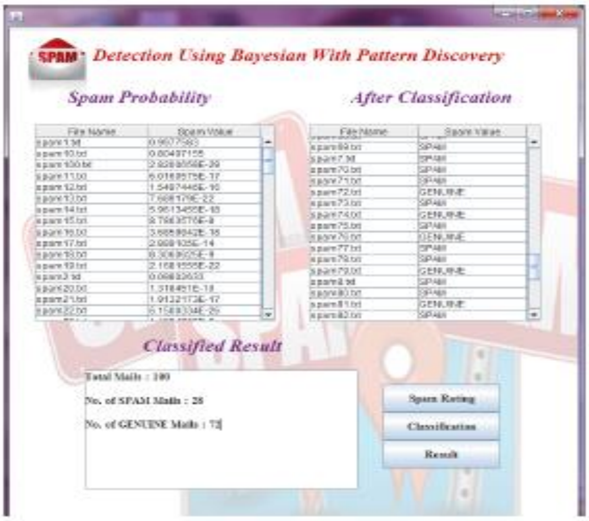
27112E-30

THIS IS A GENUINE MAIL!

Figure 10: Spam Detection for Individual Mail

V. CONCLUSION

Many data mining methods are proposed in the last decade. The proposed approach achieves and outstanding performance by text mining by comparing with other methods



File Name	Spam Value
spam1.txt	0.9577583
spam10.txt	0.80497155
spam100.txt	2.9289556E-39
spam11.txt	6.0169575E-17
spam12.txt	1.5487446E-16
spam13.txt	7.668176E-22
spam14.txt	5.9613455E-18
spam15.txt	8.7663576E-8
spam16.txt	3.6858942E-16
spam17.txt	2.969105E-14
spam18.txt	8.3069625E-9
spam19.txt	2.1681555E-22
spam2.txt	0.06832633
spam20.txt	1.318451E-18
spam21.txt	1.9122172E-17
spam22.txt	6.1583334E-26

File Name	Spam Value
spam69.txt	SPAM
spam7.txt	SPAM
spam70.txt	SPAM
spam71.txt	SPAM
spam72.txt	GENUINE
spam73.txt	SPAM
spam74.txt	GENUINE
spam75.txt	SPAM
spam76.txt	GENUINE
spam77.txt	SPAM
spam78.txt	SPAM
spam79.txt	GENUINE
spam8.txt	SPAM
spam80.txt	SPAM
spam81.txt	GENUINE
spam82.txt	SPAM

Total Mails : 100

No. of SPAM Mails : 28

No. of GENUINE Mails : 72

Figure 11: Spam Classification

In previous techniques of data mining using Discovered knowledge in the field of text mining is difficult and ineffective. Reason is that some useful long pattern with high specificity lack in support. All frequent short patterns are useful hence misinterpretation of patterns derived from data mining techniques lead to the ineffective performance. Effective pattern mining technique has been proposed to overcome the low frequency and misinterpretation problems for text mining. Proposed method uses two methods one is pattern discovery and another is the baysian classification. So we get the effective patterns using effective pattern discovery and then we use that only for the classification so we get better results than any other classifier used for the spam mail detection.

REFERENCE

1. Yuefeng Li, Abdulmohsen Algarni, Ning Zong ,âA.Mining Positive and Negative Patterns for Relevance Feature DiscoveryâAI , ACM Tranaction, KDDâAZ10 , July 2010.
2. K. Aas and L. Eikvil, âA.Text Categorisation: A Survey,âAI Technical Report Raport NR 941, Norwegian Computing Center, 1999.
3. R. Agrawal and R. Srikant, âA.Fast Algorithms for Mining Association Rules in Large Databases,âAI Proc. 20th IntâAZl Conf. Very Large Data Bases (VLDB âAZ94), pp. 478-499, 1994.
4. H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, âA.Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,âAI Proc. IEEE IntâAZl Forum on Research and Technology Advances in Digital Libraries (ADL âAZ98), pp. 2-11, 1998.
5. Y. Huang and S. Lin, âA.Mining Sequential Patterns Using Graph Search Techniques,âAI Proc. 27th Ann. IntâAZl Computer Software and Applications Conf., pp. 4-9, 2003.
6. T. Joachims, âA.A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization, âAI Proc. 14th IntâAZl Conf. Machine Learning (ICML âAZ97), pp. 143-151, 1997.
7. T. Joachims, âA.Text Categorization with Support Vector Machines: Learning with Many Relevant Features,âAI Proc. European Conf. Machine Learning (ICML âAZ98),, pp. 137-142, 1998.
8. W. Lam, M.E. Ruiz, and P. Srinivasan, âA.Automatic Text Categorization and Its Application to Text Retrieval,âAI IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
9. D.D. Lewis, âA.Feature Selection and Feature Extraction for Text Categorization,âAI Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
10. Y. Li and N. Zhong, âA.Mining Ontology for Automatically Acquiring Web User Information Needs,âAI IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.
11. Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, âA.A Two-Stage Text Mining Model for Information Filtering,âAI Proc. ACM 17th Conf. Information and Knowledge Management (CIKM âAZ08), pp. 1023-1032, 2008.
12. I. Moulinier, G. Raskinis, and J. Ganascia, âA.- Text Categorization: A Symbolic Approach,âAI Proc. Fifth Ann. Symp. Document Analysis and Information Retrieval (SDAIR), pp. 87-99, 1996.
13. J.S. Park, M.S. Chen, and P.S. Yu, âA.An Effective Hash-Based Algorithm for Mining Association Rules,âAI Proc. ACM SIGMOD IntâAZl Conf. Management of Data (SIGMOD âAZ95), pp. 175-186, 1995.
14. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, âA.Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth,âAI Proc. 17th IntâAZl Conf. Data Eng. (ICDE âAZ01), pp. 215-224, 2001.
15. M.F. Porter, âA.An Algorithm for Suffix Stripping, âAI Program, vol. 14, no. 3, pp. 130-137, 1980.
16. S. Robertson and I. Soboroff, âA.The Trec 2002 Filtering Track Report,âAI TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING. ps.gz.
17. J. Rocchio, Relevance Feedback in Information Retrieval. chapter 14, Prentice-Hall, pp. 313-323, 1971.
18. G. Salton and C. Buckley, âA.Term-Weighting Approaches in Automatic Text Retrieval,âAI Information Processing and Management: An IntâAZl J., vol. 24, no. 5, pp. 513-523, 1988.
19. S. Scott and S. Matwin, âA.Feature Engineering for Text Classification,âAI Proc. 16th IntâAZl Conf. Machine Learning (ICML âAZ99), pp. 379- 388, 1999.
20. F. Sebastiani, âA.Machine Learning in Automated Text Categorization,âAI ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
21. M. Seno and G. Karypis, âA.Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint,âAI Proc. IEEE Second IntâAZl Conf. Data Mining (ICDM âAZ02), pp. 418-425, 2002.
22. K. Sparck Jones, S. Walker, and S.E. Robertson, âA.A Probabilistic Model of Information Retrieval: Development and Comparative ExperimentsâATPart 2,âAI Information Processing and Management, vol. 36, no. 6, pp. 809-840, 2000.
23. R. Srikant and R. Agrawal, âA.Mining Generalized Association Rules,âAI Proc. 21th IntâAZl Conf. Very Large Data Bases (VLDB âAZ95), pp. 407-419, 1995.

24. S.-T. Wu, Y. Li, and Y. Xu, âA.Deploying Approaches for Pattern Refinement in Text Mining, âAI Proc. IEEE Sixth IntâAZl Conf. Data Mining (ICDM âAZ06), pp. 1157-1161, 2006.
25. S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, âA.Automatic Pattern- Taxonomy Extraction for Web Mining,âAI Proc. IEEE/WIC/ACM IntâAZl Conf. Web Intelligence (WI âAZ04), pp. 242-248, 2004.
26. X. Yan, J. Han, and R. Afshar, âA.Clospan: Mining Closed Sequential Patterns in Large Datasets,âAI Proc. SIAM IntâAZl Conf. Data Mining (SDM âAZ03), pp. 166-177, 2003.
27. M. Zaki, âA.Spade: An Efficient Algorithm for Mining Frequent Sequences,âAI Machine Learning, vol. 40, pp. 31-60, 2001.