Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm

Kavitha Murugeshan, Neeraj RK

Abstract— Text mining has been an unavoidable data mining technique. There are different methods for text mining, One of the most successful will be mining using the effective patterns. Here a Naïve Bayesian algorithm is being used for discovering of patterns, since this will be the most appropriate one for classifying positive and negative documents. The usual results will not be in an optimized manner. The prescribed method makes the output arranged in a particular order.

Index Terms— Text mining, Pattern discovery, Pattern Taxonomy

I. INTRODUCTION

As the World is being improvised in a digital way, knowledge discovery and Data mining have an important task. Useful information is always needed in all sort of information extraction. Text mining is therefore a step in knowledge discovery process in Databases and Datasets

Many data mining techniques have been proposed for mining useful patterns 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 existing, Information Retrieval (IR) provided many term-based methods to solve this challenge. The term-based methods suffer from the problems of polysemy and synonymy. Polysemy stands for a word having different meanings, and synonymy stands for different words having the same meaning.

The proposed paper we use pattern (or phrase)-based approaches which perform better in comparison studies than other term-based methods. This approach improves the accuracy of evaluating support, term weights because discovered patterns are more specific than whole documents.

II. RELATED WORKS

Here we are proposing a pattern taxonomy model. Other different pattern mining methods are Sequential patterns, Sequential closed patterns, frequent itemsets,Frequent closed item sets. All these provide similar results but on depending on precision and recall our method stand way apart. The curve for PTM will remaining better and smoother when compared to the other pattern mining methods. When recall value raises the pattern mining methods starts coming down abruptly. This is shown in the figure.

Manuscript received May, 2013.

Mrs.Kavitha Murugeshan , Computer science and engineering, Calicut University/VVIT/Karadparamba,Malappuram,India.

Mr.Neeraj.RK,Computer science and engineering,Calicut University/ VVIT/ Karadparamba,Malappuram,India.

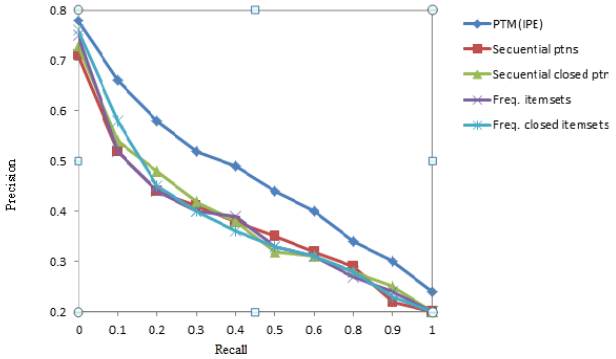

Fig.1: When recall value raises the pattern mining methods starts coming down abruptly

III. PATTERN TAXONOMY MODEL

As a first step in this paper the given documents are separated into different paragraphs. So consider every document d generates a set of paragraphs say, PS(d). Assume D is a set of documents, Which consists of two sets. A set of positive documents, D+; and a set of negative documents, D-.Let T ={t1; t2; . . . ; tm} be a set of terms (or keywords) which can be extracted from the set of positive documents, D+.

Positive documents are the documents that are that are frequently occurring .Here we are only considering the positive documents**. For the classification into positive documents we are using the Naïve Bayesian Algorithm**.

IV. CLASSIFICATION

The various data in a classification is related with different cases or categories. Classification is related to handling the data functions. it assigns items in a collection to various targets target categories or classes. Any classification task primarily begins with a data set which contains the various known class assignments. The aim of classification is to precisely predict the target class for each case in the data items. Consider a classification model which could be used to identify loan applicants as low, medium, or high credit risk people. The classification task begins with the collection of various data sets that affect the credit risks, in the sited example; the class assignment is the credit risk and the data sets associated with the credit risks are associated as follows. The credit risk can be majorly developed based on historically observed data of many loan applicants over a time period. Besides the data like history of employment, whether home is owned or rental, number of years of residence , and so on attribute to the credit risking of the many loan applicants. It's clear that the target is credit rating and the other attributes would be predictors only; and the data for each customer would constitute a case. The target can be numerical or

categorical .The numerical target can be floating point values. Classifications are not continuous and may not have any order. Binary classification is the simplest type of classification problem. In binary classification, the target attribute has the values either high or low: for example, credit rating with high or low values. Multiclass targets are having more than two values: for example credit rating may be of high low, medium or unknown. Classification algorithms find relationship between the values of predictors and values of target. Different Classification algorithms use different techniques for finding relationships.

A model is used to conclude the obtained relationships, which can then be applied to a different data set in which the class assignments are unknown. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. The data set for a classification system divided into two: one for building the model; the other for testing the model. Classification models are tested by comparing the predicted values to known target values in a set of test data. The probability for each case is also decided by scoring a classification model; for example model that classifies customers as low, medium, or high value credential risks would also predict the probability of each classification for each customer. In this work, we classify terms as positive and negative using Naive Bayes Classifier.

A naive Bayes classifier

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. The model based on this classifier would be more precisely called as an "independent feature model". We are using weka tool for classifying the terms in Naïve Bayes Classifier .this classifier builds an independent feature model. For example, a fruit may be considered to be an apple if it has the following features i.e. if it is red, 4" in diameter and round. This classifier considers all these features to contribute independently to the probability that the fruit is an apple, i.e. these features can be related to each other or to the existence of the other features. Thus Naïve Bayes classifier acts as a realistic probabilistic classifier. It can be also used in supervised learning setting the method of maximum likelihood is used in many applications and certain estimations for naive Bayes models. Therefore its easy to work with the naive Bayes model even if one don't believe in Bayesian probability or any Bayesian methods.

V. D-PATTERN MINING AND INNER PATTERN EVOLUTION

To improve the efficiency of the pattern taxonomy mining, an algorithm, SP Mining, was proposed to find all closed sequential patterns, which uses Apriori property in order to reduce the searching space. Algorithm shown describes the training process of finding the set of d-patterns. Positive documents are found using naive Bayesian classifier after that the SPMining algorithm is first called in step 4 giving rise to a set of closed sequential patterns SP. The paper consists of the d-pattern discovery and term support evaluation which comes under deploying process. In Algorithm all discovered patterns in a positive document are composed into a dpattern giving rise to a set of d-patterns DP in steps 6 to9. From steps 12 to 19, term supports are calculated based on the normal forms for all terms in dpatterns.

Here an equation is used for the calculation of term weight

$$\text{Weight}(D) = \sum \text{Support}(t)\,,\tau(t,d) \qquad (1)$$

We have support(t) defined in Algorithm and

$\tau(t,d) = 1, \tau \epsilon\, d$; otherwise $\tau(t,d) = 0$

Input: *positive documents D^+; minimum support, min_sup.*

Output*: d-patterns DP and support of terms*.

1: *DP* = $\emptyset$;

2: foreach *document d* $\in D^+$ **do**

3: *let PS(d) be the set of paragraphs in d;*

4: *SP* = SPMining*(PS (d), min_sup);*

5: $\hat{d} = \emptyset$;

6: foreach *pattern* $p_i \in SP$ **do**

7: $p = \{(t, 1) | t \in p_i\}$;

8: $\hat{d} = \hat{d} \oplus p$;

9: end

10: $DP = DP \cup \{\hat{d}\}$;

11: end

12: $T = \{(t | (t, f) \in p, p \in DP\}$;

13: foreach *term* $t \in T$ **do**

14: *support(t)* =0;

15: end

16: foreach *d pattern* $p \in DP$ **do**

17: foreach $(t,w) \in \beta(p)$ **do**

18: *support(t)= support (t) + w;*

19: end

20: end

Algorithm 1 : D-pattern mining Algorithm

"Classification is a classic data mining technic based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming and statistics. in data mining Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes"

Example 1.

Consider this paragraph as an example. when we applied the above algorithm the result formed was like this

The set of D-patterns = {mine(t1) ,technic(t2),item(t4)}

When pattern deployment Algorithm was applied we got three patterns from paragraph1 (dp1). The three corresponding Patterns were

No.	Pattern.	Weight.	Support
1	{ti,t2}	1	3
2	{t1,t4}	1	2
3	{t1,t2,t4}	1	2

Table.1:Table of Three Corresponding Patterns

Inner Pattern Evolution

In this section, we discuss about the shuffling of supports of terms d-patterns based on negative documents in the training set. This reduces the side effects of noisy patterns because of the low-frequency problem. It changes only a pattern's term supports within the pattern, this technique is called inner pattern evolution. Documents into relevant

or irrelevant categories based on a Threshold. The main process of inner pattern evolution is implemented by the algorithm IPEvolving (see Algorithm 2). The inputs of this algorithm are a set of d-patterns DP, a training set D=D+ Ụ D-. The output is a composed of d-patterns. Step 2 estimates the threshold for finding the noise negative documents. Thereafter Steps 3 to 10 go over term supports by using all noise negative documents. Step 4 is for finding the noise documents. Step 5 gets normal forms of dpatterns NDP. Step 6 calls algorithm Shuffling (see Algorithm 3 in) to update NDP according to noise documents. Thereafter Steps 7 to 9 binds the updated normal forms together.

Input: *A training set D=D+U D-,a set of D patterns DP and an experimental co efficient μ*

Output: *A set of term support pairs np*

1.Np ←φ
2.Threshold= Threshold(DP);//
3.**Foreach** noise negative document nd Є D- **do**
4.If weight(nd) ≥ threshold **then**,Δ(nd)={p element DP |termset (p) ∩ nd ≠φ};
5.NDP={β(p)|p Є DP};
6.Shuffling(nd, ,Δ(nd), NDP,μ NDP);// call Alg 3:
7.**foreach** p Є NDP do;
8.np ← np Θ p;
9.end
10. end

Algorithm 2 : IPEvolving(D+,D-,DP,μ)

In algorithm 3The parameter offering is used in step 4 for the purpose of storing the reduced supports temporarily of some terms in a partial conflict offender. Here the offering is part of the sum of supports of terms in a d-pattern where these terms also appear in a noise document. Thereafter the algorithm calculates the base in step 5 which is non-zero .The updation of the support distributions of terms is done in step 6.

Input: *A noise document nd ; its offendersΔ (nd); normal patterns of D patterns NDP and an experimental coefficient μ.*

Output: *normal forms of d- patterns NDP which is updated.*

1: foreach *d-pattern p in* Δ(*nd*) *do*
2: if *term set (p)* ⊆ nd; **then** *NDP= NDP-{β (p)};*
3: else *partial conflict offenders*
4: *Offering* = $(1-\frac{1}{\mu}) \times \sum_{t \in (termset(P) \cap nd)} support(t)$;
5: *base* = $\sum_{t \in (termset(P) - nd)} support(t)$;
6: foreach *term t in termset (p) term t* **do**
7: if $t \in nd$ **then** $support(t) = \frac{1}{\mu} \times support(t)$;// *shrink*
8: else // *grow supports*
9: $support(t) = support(t) \times (1 + offering \div base)$;
10:end
11: end

Algorithm 3 : Shuffling(nd,Δ(nd),NDP,μ.NDP)

On considering example 1 t1,t2,t4 are detected as the noise patterns . since the weight of each pattern is 1 the threshold is calculated to be 0.Thereafter t1,t2 and t4 are shuffled using the algorithm3. And we obtained same support and total weight for all the three terms.

The proposed model includes two phases 1)the training and 2) the testing . In the first phase, the proposed model first calls Algorithm PTM (D+,min_sup) to find d-patterns in positive documents (D+) based on a min_sup, and evaluates term supports by deploying dpatterns to terms. It also calls Algorithm IPEvolving (D+,D-,DP,μ)to revise term supports using noise negative documents in D- based on an experimental coefficient μ . In the testing phase, it evaluates weights for all incoming documents using eq. (1). The incoming documents then can be sorted based on these weights.

VI. CONCLUSION

The research area is an emerging technology. Search engines could use this up to prevent the black optimizers. Thus doing research in this area explores new area of study with more scope. Many data mining techniques have been proposed in the last decade including association rule mining, mining using frequent itemsets, mining of sequential patterns and closed patterns, maximum pattern mining. However, using these patterns in the field of text mining is difficult and ineffective. This is because some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems .Here we uses two processes, pattern deploying and pattern evolving, which refines the discovered patterns in text documents. And also use Naïve bays classification approach to classify the terms.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. "Effective Pattern Discovery" .2012-IEEE
[2] K. Aas and L. Eikvil, "Text Categorization: A Survey," Technical Report Report NR 941, Norwegian Computing Center, 1999.
[3] 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.
[4] 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.
[5] R. Baeza-Yates and B. Ribeiro-Neto,Modern Information Retrieval. Addison Wesley, 1999.
[6] 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.
[7] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "WordSequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.
[8] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrasesin Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell'Informazione, 2000.
[9] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[10] S.T. Dumais, "Improving the Retrieval of Information fromExternal Sources," Behavior Research Methods, Instruments, andComputers, vol. 23, no. 2, pp. 229-236, 1991.
[11] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence,"Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[12] 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.
[13] Y. Huang and S. Lin, "Mining Sequential Patterns Using GraphSearch Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.

Mrs.Kavitha Murugeshan is the head of the Computer Science department in Vedavyasa Institute of Technolgy under Calicut university.She is persuing her Phd in Datamining.she did her ME in Anna niversity.

Mr.Neeraj Rk is pursuing Mtech from Vedavyasa Institute of Technology under Calicut university. He completed his Btech degree from Calicut University