Classification of Documents using Effective Pattern Taxonomy

Mallareddy Uday Kiran
CSE, MVGR College of Engineering
Vizianagaram, AP

R Ravikanth
Assistant Professor
CSE, MVGR College of Engineering
Vizianagaram, AP

ABSTRACT
Text mining is a technique helps users in extracting useful information from large amount of database available digitally on web or text data. Pattern Taxonomy based model containing sequential pattern used to perform the task. EPT (Effective Pattern Taxonomy) method helps in extracting useful patterns in the text documents by classifying them in Positive and Negative documents. Pattern-based method outperforms keyword based methods. Pattern based method is best way to eliminate meaningless as well as closed sequential patterns thus saving computational time and increases effectiveness of the system.

KEYWORDS
Pattern mining, Positive/Negative documents, Effective Pattern Taxonomy, Sequential patterns, Precision/Recall

1. INTRODUCTION
Data mining is the process of extracting useful information and knowledge from the data. Data mining also referred as the knowledge mining from the data, extraction of the knowledge, pattern/data analysis, data dredging & data archeology. Knowledge discovery from data is also synonymy for data mining.

Data mining is an interesting process where many useful and intelligent methods are applied to extract data patterns from the data. The scope of the data mining tends to data repositories such as files, data streams, transactional databases, world wide web, documents. There are many mining techniques such as Data streams mining, Time-series data, Graph mining, Link mining, multi relational mining, spatial data mining, Multimedia data mining & Text mining.

In data mining a substantial portion of available information is stored in text databases which consists of large collection of documents from various forms such as articles, books, e-mail messages, documents & various kind of electronic documents Text mining is the process of knowledge discovery in text documents.

Extracting useful knowledge in text documents is done by many methods such as term based methods, Phrase based methods & pattern mining based methods.

Information Retrieval(IR) is an text retrieval method of extracting text. IR includes an term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning.

So Phrases based approach has been carried out in order to overcome Term based approach but it suffered from the problem of low frequency of occurrence and redundant as well as noisy patterns among the phrases. The semantic meaning of many discovered terms is uncertain for answering what users want.

Pattern mining helps in finding the specific patterns in text documents. By evaluating weights feature in text document pattern mining is carried out. To overcome above hypothesis pattern based approach is carried out which helps in identifying important patterns in the text document. Pattern mining uses the concept of closed sequential patterns and pruned nonclosed patterns.

In this Paper it is proposed Pattern mining by categorizing patterns by Sequential pattern in Positive and Negative documents which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem.

The work below is categorized in the sections where in Section 1 it gives the introduction while in Section 2 it is sorted out literature work of all the experts in field of text mining and pattern mining and coming on to Section 3 the use of Pattern Taxonomy model application to carry out task while in the Section 4 there is algorithm which gives information of the work in this paper while Section 5 and Section 6 is about experimental results and conclusions and there is references of all the papers research work done by various members in this field.

2. RELATED WORK
The literature consists of effective work done for the various mining techniques available which are as follows.

Ning Zhong, Yuefeng Li, and Sheng-Tang Wu addresses that term based method suffers from the problems of polysemy and synonmy and they suggested that Pattern based method performs better than term based methods. For finding relevant information they have used processes of Pattern Deploying and Pattern Evolving.

Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen have discovered valuable topics from Text sequences by the two Distibutions: Topics intensity over time by Time distribution & Semantic of the Topic by word Distribution.

Yuefeng Li and Ning Zhong's approach is to discover ontologies from the data.he has proposed ontology mining algorithm to differentiate between the documents whether it is positive or negative. They have presented a novel technique to capture patterns in ontology.

Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE proposed concept based mining technique which calculates similarities in documents by matching concepts between documents and also by semantics of sentences.

Nikky Rai, susheel jain, Anurag jain used the concept of association rule to evaluate differences between Positive and negative patterns in the documents.They have proposed Association rule mining algorithm to extract useful patterns from the large data.

Seema Mishra, G C Nandi have presented a novel frequent pattern mining based approachto identify frequent person detection i.e. apriori to solve frequent association problem between social networks obtained from low level task of face recognition.

3. PATTERN TAXONOMY MODEL:

Pattern Taxonomy is an new pattern-based model for representing Text documents. It is a Tree-like Structure which depicts out patterns being extracted from a text data.

Instead of the keyword-based concept used in thetraditional document representation model, the pattern based model containing frequent sequential patterns (single term or multiple terms) is used to perform the same concept of task.

Here all the documents are being splitted into paragraphs. All the documents consists of positive (D+) and negative (D-) documents. Here categorize the documents as you like. For Example, All the Technology related documents can be termed as Positive/Negative as well as all the Science related documents can be termed as Positive/Negative. Let T be the terms to be extracted from set of Positive/Negative documents.

We discover a sequential pattern form collection of text documents and generate this pattern taxonomy model to depict relation between patterns being observed in the documents.

Let us consider a small example for Pattern Taxonomy:

Given Sequential Pattern P= (I1, I2, I3) in database.

Lets have minimum support min_sup=75%. Also given is the Paragraph Ids where this pattern occurs.

Paragraph ID	Sequence
1	<I1, I2, I3, I4>
2	<I2, I4, I5, I3>
3	<I3, I6, I1>
4	<I5, I1, I2, I7, I3>

Table 1. Sequences in the document

Basing on the minimum support 4 patterns arrived <I2,I3>, <I1>, <I2>, <I3> .

The purpose of minimum support min_sup to reduce number of patterns discovered in the document. Else pattern with lower relative support increases burden of training so, removing less significant patterns will certainly save the time.

Now sort out sequences and as well as their subsequences.

Example 1: For Example pattern <I1, I2> is a subsequence of <I1, I2, I3> and pattern <I2> is a subsequence of <I2, I3> then you can identify that pattern <I1, I2> is already subsequence of <I1, I2, I3> so instead of using <I1, I2> you can use long pattern <I1, I2, I3> Once you construct the tree then find relevant patterns and remove irrelevant patterns.

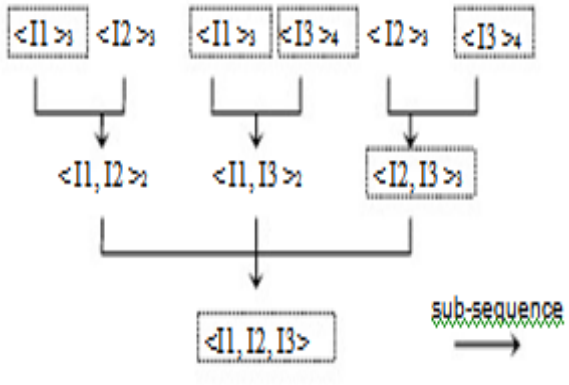

Figure 1. Pattern taxonomy for sample data

As you can see in the above diagram that <I1, I2> is an closed pattern to the pattern <I1, I2, I3> that means they always appear in the document . So the short pattern <I1, I2> is negligible as it is meaningless since the long pattern carry more information than shorter one.
In order to remove such patterns use pruning procedure where unnecessary and closed patterns are removed. The main aim is to eliminate meaningless patterns.
Let us observe a small example:

Example 2. As you see in the Figure 2 you see that the super pattern is <I1, I2, I3> and the closed patterns are <I2>, <I4>, <I5>, <I1,I2>, and <I1, I3> which are pruned . For a look pattern <I2> has four of the super patterns <I1, I2>, <I2, I3>, <I2, I4> and <I1, I2, I3> whose support is same as the pattern <B>'s is. The dash line in the figure indicates that linked patterns have close relation- to their linked patterns.

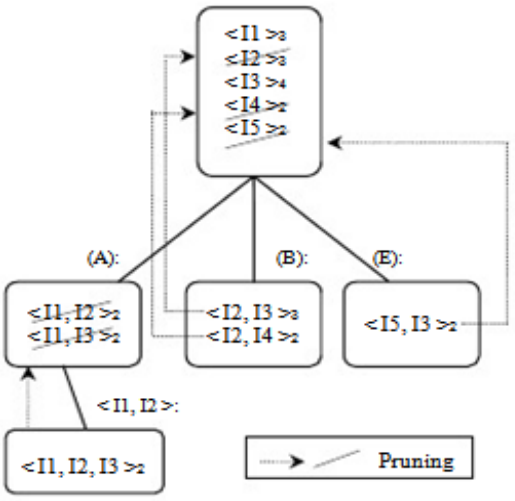

Figure 2. Pruned pattern

3.1 Method for Deploying pattern

To deploy a pattern there is need to identify the terms in the document and depict the pattern where all the terms are forming in the document.

A set of terms referred as termset .In a document d presume the terms in the document as t and define the number of occurrence of term t in document d as tf(t,d) . Pattern can be refferd as P = {(t,f) | t ϵ T, f = tf(t,d) > 0}, where T is referred as stack of set of keywords such as terms. You also use support(P) because the greater the support is the more important the pattern is. Let termset (P) = {t|(t,f) t ϵ P} and in this paper if termset(P1) = termset(P2) the P1=P2 .Termset defines uniqueness in pattern. You can compose two patterns by using composition operator iff termsets in both the patterns ar equal. Let P1 and P2 be two patterns. Call P1*P2 be the composition of P1 and P2 where it satisfies:

P1*P2 = {(t, f1 + f2) | (t, f1) ϵ P1, (t,f2) ϵ P2} {(t, f) | t ϵ (termset(P1) *termset(P2)) – (termset(P1) *termset(P2)), (t,f) ϵ P1*P2}
Support (P1*P2) = support (P1) + support (P2).

So using the above concept let us have an example:
Example 3: For example, P1 = <t1, t2> and P2 = <t2, t3>. The termsets of P1 and P2 are {(t1, 1), (t3,1)} and {(t2,1), (t3, 1)} respectively. Composition of two patterns can be depicted as P3 where termsets are termset(P3) = {(t1, 1), (t2, 1) (t3, 1)}.

3.2 Application

The Pattern Taxonomy model method applied to both the Positive and Negative documents by which you get required knowledge from training set of user profiles.

Remove stop word and redundant words as well as subsequent words which occurs many times and has less meaning.

There are two phases in this application phase i.e. Training and Testing.

Training phase used to find all frequent patterns from entire documents and prunes meaningless words and then calculates Weights of all patterns. A centroid is used to hold the representation of patterns extracted from training set. The value of Pattern P is computed by following weight function:

W(P)= (|{da| da∈ Dpos, P in da }|)/(|{db | db ∈ Dneg, P in db }|)

da and db denotes documents,

D denotes training set

Dpos denotes positive documents

Dneg denotes negative documents

where as Dpos and Dneg are the subsets of D.

Once centroid is obtained then the testing phase is carried out.

Testing phase finds differences in positive/negative documents by the centroid obtained in training phase by ranking each of them. The simple way to estimate similarity between documents and centroid by summing weights of patterns which are in the documents.

In the same fashion weights of patterns is calculated you can also find weights of individual terms in the document by a simple formula

W (t)=log⁡〖 (d/(D-d))/((n-d)/((N-n)- (D-d)))〗

N is the total number of documents in training set,
D is the number of relevant documents,
n is the number of documents which contains t,
d is the number of relevant documents which contains t.

4. ALGORITHM

Train:--
Step1: Taking positive(Technology related document) and negative(Science related document) documents to train
Step2:
Begin
b1= positive document
b2= negative document
if(b1== Tech) //* Tech = Technology document
print("click on positive document to choose");
//select folder for positive document to train
else
if(b2== Sci) //* Sci = Science document
print("click on negative document to choose");
//select folder for negative document to train
Step3: Now train both the documents
Step4:
if(b1== b2)
print("Select different path to choose positive and negative documents");
else
// perform Pattern Taxonomy Method to find out frequent pattern and weights of the individual patterns as well as terms(see Section 3)

Test:--
Step5: Testing accuracy among the documents which you choose
Step6: Select both documents to compare accuracy
if(b1==b2)
//Result is false positive since you have selected positive and negative documents for inputs
else
//result is true positive and result is accurate

Prediction:--
Step7: Take a document as input
if
//Taken file is related to technology it classifies the document as positive document
else
//It is negative document

End

5. EXPERIMENTAL RESULTS

You now evaluate results for proposed approach Effective Pattern Taxonomy Method and Deploying method .To determine accurate measures of similarity or difference between documents you depict results by graph pattern and table pattern.

The experimental setup consists of relevant documents that you termed as positive and negative documents .i.e Technology related (Positive) and Science related (Negative). You take into account support factor also consisting of support = 50, support = 75 and support= 90. You also take training set of 6 documents in which you take 3 positive

documents and 3 negative documents. Basing on the support factors you have calculated documents weights and also uniquely calculated document weights by PTM method which you discussed earlier.

Table 2. Support factors for different documents

Doc/Sup	Sup =50	Sup =75	Sup =90
Doc 0	1.31556	1.80357	1.80357
Doc 1	1.75182	1.87012	1.87012
Doc 2	2.13338	2.27178	2.27178

In the above Table 2 you have 3 documents with there relevant weights basing on the support factor.

You now compute the above table in pictorial pattern which helps in easy way of understanding.

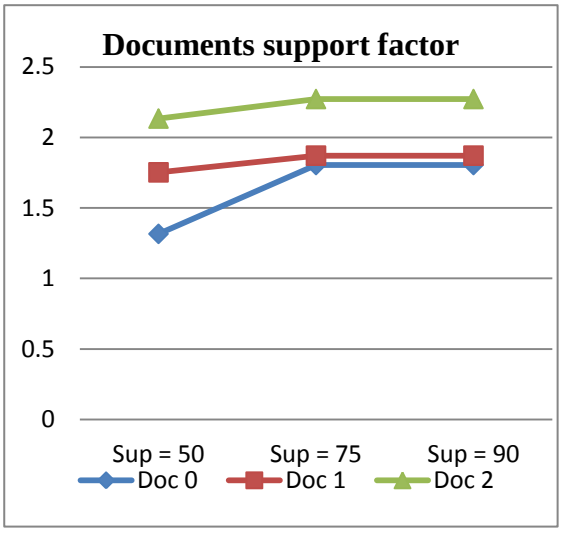

Figure 3. Document/support factors pictorial representation by weight factor

In the above Table 4 you see that documents are categorized in positive/negative documents. So by using PTM categorize the document and then weights are being calculated for individual patterns. So if you see whenever the support factor is being increased you see weights are being increased. That means when ever support is increased then the relevance computation and performance of all the patterns in the documents is increased and results are more effective.

Let us have an Pictorial representation of the above discussion in the below Figure 4.

The green color line in the figure depicts for the support factor = 90 for positive and Negative documents. The Red color computes support = 75 for the documents that gives as input.The blue color indicates for the support = 50.

Table 3.Weight factors of various positive/negative documents

Doc/sup	Weights	Sup = 50	Sup = 75	Sup = 90
Doc 1	Pos doc	1.31556	1.80357	1.80357
	Neg doc	0.56941	1.0743	1.0743
Doc 2	Pos doc	1.75182	1.87012	1.87012
	Neg doc	0.51786	0.9243	0.9243
Doc 3	Pos doc	2.13338	2.27178	2.27178
	Neg doc	2.13338	1.16655	1.16655

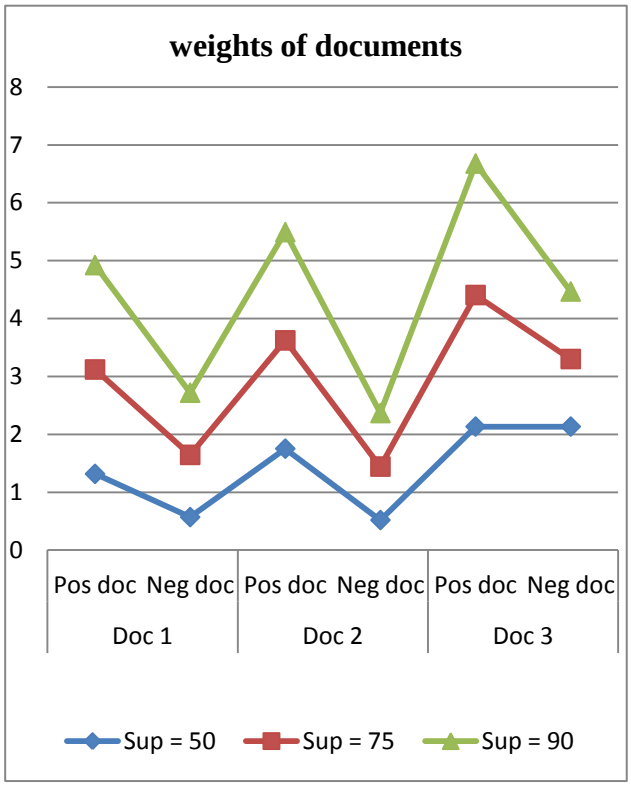

Figure 4. Weights of positive/negative documents

By basing on the above figure it is clear that by increasing the Support factor there is effective results of getting patterns. Finally you achieve better performance.

6. CONCLUSION

Many Text mining techniques have been used in the last decade. The techniques included there are frequent itemset mining, sequential pattern mining, closed pattern mining. However effectiveness of text mining system has not been improved very much. Phrase based method is also available which gives inconsistent results so in this paper use of the pattern based approach by using the support factors effectively. You have pruned meaningless patterns and also removed stem words and stop words. This paper can be extended by dividing positive/negative documents basing on the patterns in the documents. Also weight factor can be optimized in the documents pattern. This paper can be extended to the scope of Web based mining. Document separation is the key factor in our document.

7. REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining" in IEEE transaction, vol. 24, January 2012.

[2] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE "An Efficient Concept-Based Mining Model for Enhancing Text Clustering" IEEE transactions on knowledge and data engineering, vol. 22, no. 10, October 2010.

[3] Kavitha Murugeshan, Neeraj RK "Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm" IJITEE ISSN: 2278-3075, Volume-2, Issue-6, May 2013.

[4] Yuefeng Li Abdulmohsen Algarni Ning Zhong "Mining Positive and Negative Patterns for Relevance Feature Discovery".

[5] Nikky Rai, Susheel Jain, Anurag Jain "Mining Positive And Negative Association Rule From Frequent And Infrequent Pattern Based On Imlms_Ga" IJCA (0975 – 8887)

[6] Abdulmohsen Algarni, Yuefeng Li, Xiaohui Tao,"Mining Specific and General Features in Both Positive and Negative Relevance Feedback".

[7] T. Joachims. "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization". In Proc. Of ICML'97, pages 143–151, 1997.

[8] S. Shehata, F. Karray, and M. Kamel. "A concept-based model for enhancing text categorization". In Proc. Of KDD'07, pages 629–637, 2007.

[9] 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.

[10] S. Scott and S. Matwin, "Feature Engineering for Text Classification," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 379- 388, 1999.

[11] "Automatic Pattern-Taxonomy Extraction for Web Mining" Sheng-Tang Wu Yuefeng Li Yue Xu Binh Pham Phoebe Chen* IEEE Conference.

[12] R. Agrawal, and R.Srikant, "Mining sequential patterns," Proceedings of Int. Conf. on Data engineering (ICDE'95), Taipei, Taiwan, 1995, pp. 3-14.

[13] G. Chang, M.J. Healey, J. A. M. McHugh, and J.T. L. Wang, "Mining the World Wide Web: an information search approach", Kluwer Academic

Publishers, 2001, pp. 192.

[14] D. A. Grossman and O. Frieder, "Information retrieval algorithms and heuristics", Kluwer Academic publishers, Boston, 1998.

[15] J. D. Holt and S. M. Chung, "Multipass algorithms for mining association rules in text databases", Knowledge and Information Systems vol.3, 2001, pp. 168-183.

[16] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," ICDM03, 2003, pp. 179- 186.