Classification of News and Research Articles Using Text Pattern Mining

Sujit V Chaudhari Shrikant Lade
RKDF IST

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 approaches are better than the term based approaches. But pattern based approach is better than the term based and phrase based approach. The proposed method utilize pattern approach with the set of keywords, which is an innovative and effective pattern discovery technique by which research articles, news articles classification of different field are done and more than 80 percent of documents are successfully identified.

Index Terms
Computer Networks, Network Security, Anomaly Detection, Intrusion Detection.

1. INTRODUCTION

Text Mining [1] is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems Undeterred by the text explosion. It involves analyzing a large Collection of documents to discover previously unknown Information. The information might be relationships or Patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the Interest at present is coming from the biological sciences. Originally, research in text categorization addressed the binary problem,
Where a document is either relevant or not. Text mining involves the application of techniques from areas such as information retrieval, natural language Processing, information extraction and data mining.

Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The most well-known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. This is however changing with the advent of digital libraries, where the documents being retrieved are digital versions of books and journals. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

2. LITERATURE VIEW

Many types of text representations have been proposed in the past. A well-known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting. [1], the combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase-based text representation for Web document management was also

Manuscript received October 5, 2015
Manuscript revised October 20, 2015

proposed in [2].In [3]; data mining techniques have been used for text analysis by extracting co-occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned in [4]. In, hierarchical clustering [5], [6] was used to determine synonymy and hyponymy relations between keywords. Pattern mining has been extensively studied in data mining communities for many years. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [7], [8], [9]. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent item sets, co-occurring terms and multiple grams, for building up a representation with these new types of features. Nevertheless, the challenging issue is how to effectively deal with the large amount of discovered patterns. For the challenging issue, closed sequential patterns have been used for text mining in [10], which proposed that the concept of closed patterns in text mining was useful and had the potential for improving the performance of text mining. Pattern taxonomy model was also developed in [11] and [10] to improve the effectiveness by effectively using closed patterns in text mining. In addition, a two-stage model that used both term-based methods and pattern based methods was introduced in [12] to significantly improve the performance of information filtering. Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model [13], was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. Pattern based methods
was introduced in [12] to significantly improve the performance of information filtering.

3. Related Work

The feature selection involves the indexing; tokenizing the text, feature space reduction. There are mainly two approaches in the text categorization knowledge engineering approach and machine learning approach [3]. In knowledge approach the user defines the rules manually Bag of words is one of keyword based method that is widely used. Simplicity is the benefit of this approach. The extracted words from the document are stored in the feature space. Synonyms and homonyms are the disadvantage of this approach. Selecting the limited number of features and over fitting are another issue.

A. Phrase-Based Representation

Keyword representation causes the uncertainty issue. Phrases consist of more precise content than single word. It can automatically discover the hidden semantic sequences of each category in documents; this can profit the classification accuracy. N-multigram model is related to n-gram model.

B. Vector Based Method

There are two types of vector based methods centroid algorithm and support vector machine. The simplest one is centroid algorithm. New document can be easily categorized by this algorithm for each category average feature vector is calculated during learning stage. In support vector machine we use negative documents along with the positive documents. Superior runtime behavior is the advantage of this method at the time of categorizing the new documents because dot product is calculated for every new document.

C. Pattern Based Method

There are two factors concerning the efficiency of pattern-based approaches: low frequency and misinterpretation. If the minimum support is decreased noisy patterns can be found lots. Misinterpretation is the measures used in pattern mining turn out to be unsuitable in using discovered patterns to answer what users want. The difficult problem hence is how to use discovered patterns to accurately estimate the weights of useful features

Different modules of the paper are
2. A. Text Preprocessing
3. B. Pattern taxonomy process
4. C. Pattern deploying
5. D. Pattern evolving

D. Text Pre-Processing

Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification. Here we read whole project and put all words in the vector. Now again read the file which contain stop words then remove similar words from the vector. Once the data is pre-process then it will be the collection of the words that may be in the ontology list. For example let one paper of the image class is taken

and its text vector is {a1, f1, s1, a2, s2, a3, a4, f2……..an} and let the stop words collection is {a1, a2, a3 ….am}. Then the vector obtain after the Pre-Processing is {f1, s1, s2, f2 ….fx}.

E. Pattern taxonomy process

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 suppor**t 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 show documents and terms sets Table 2 show 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.

Frequent and Closed Patterns Given a termset X in document d, X' is used to denote the covering set of X for d, which includes all paragraphs dp Є PS(d) such that X subset of dp, i.e. X' = {dp|dp Є PS(d), X subset of dp}. Its absolute support is the number of occurrences of X in PS (d), that is sup(X) = X'. Its relative support is the fraction of the paragraphs that contain the pattern, that is, supr (X) =X'/Ps (d).

A termset X is called frequent pattern if its supr >=min sup, a minimum support.

Table 1.

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

Table 2.

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

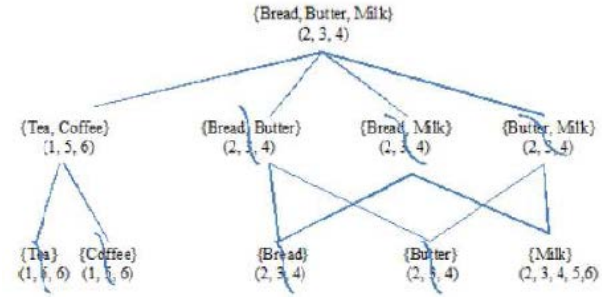
Table 1 lists a set of paragraphs for a given document d, where PS(d) = (dp1; dp2; . . . ; dp6}, and duplicate terms were removed. Let ten frequent patterns in Table 1 using the above definitions. Table 2 illustrates the ten frequent patterns and their covering sets. Not all frequent patterns in Table 2 are useful. For example, pattern {t3; t4} always occurs with term t6 in paragraphs, i.e., the shorter pattern, {t3; t4}, is always a part of the larger pattern, {t3; t4; t6}, in all of the paragraphs.

Hence, we believe that the shorter one, {t3; t4}, is a noise pattern and expect to keep the larger pattern, {t3; t4; t6}, only.

Given a termset X, its covering set X' is a subset of PS (d), we can define its termset, which satisfies

Termset (Y) = (t | for_all dp Є Y→ t Є dp)

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).



Pattern Mining Algorithm

Here this work as the training module or the pattern preparation with some minimum support. Following are the inputs to the algorithm: D training document, minimum support Ms value for the pattern List of Keywords K.

1. Ps[n]←Find_paragraph(D)

2. Loop 1:n
3. Sp[m]←Find_pattern(K, Ps, Ms)
4. Loop i =1:m
5. p ={(t,1)|t Є Sp[i]}
6. d = d + p
7. end
8. Dp = Dp U d
9. End
10. Loop p Є Sp
11. Loop (t,w)Є p
12. Supp(t) = supp(t) + w
13. End
14. End

In above algorithm

Ps: Paragraph
n: number of paragraph
Sp: Sequence Pattern
m: Number of pattern
t: term
Dp: Deploying Pattern
Supp: Support

Paragraph from each document is generate then the patterns are find by passing the keyword set K, Paragraph Ps, minimum support Ms. This will generate different sequence patterns now collect it in Sp, then find terms frequency in the paragraph. Now calculate the weight for each term, finally generate the deploying pattern. Generate closed pattern for each term in all pattern of different paragraph.

INNER PATTERN EVOLUTION

In this section, we discuss how to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. 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:

$$Threshold(DP) = \min_{p \in DP} \left(\sum_{(t,w) \in \beta(p)} support(t) \right) \quad \text{Eq. 1}$$

A noise negative document nd in D_ is a negative document that the system falsely identified as a positive, that is weightðndÞ _ ThresholdðDPÞ. In order to reduce the noise, we need to track which d-patterns have been used to give rise to such an error. We call these patterns offenders of nd.

There are two types of offenders: 1) a complete conflict offender which is a subset of nd; and 2) a partial conflict offender which contains part of terms of nd.
The basic idea of updating patterns is explained as follows: complete conflict offenders are removed from d-patterns first. For partial conflict offenders, their term supports are reshuffled in order to reduce the effects of noise documents.
For example, for the following d-pattern

d' = {(t1; 3), (t2; 3); (t3; 3) (t4; 3); (t6; 8)}

Its normal form is {(t1; 3/20); (t2; 3/20); (t3; 3/20); (t4; 3/20); (t6; 2/5)}
Assume nd = {t1; t2; t6; t9}, d' will be a partial conflict offender since

termset (d')∩ nd= {t1; t2; t6}

Let μ = 2, offering = ½(3/20+3/20+ 2/5) = 7/20 , and base = 3/20 + 3/20 =3/10 . Hence, we can get the following updated normal form by using algorithm Shuffling:
{(t1; 3/40); (t2; 3/40); (t3; 13/40); (t4; 13/40); (t6; 1/5)}
Algorithm for Inner Pattern Evolution

1. Tr←Threshold(Dp)
2. Np[n]←Find_pattern(Dn)
3. Loop i = 1:n
4. If weigth(Np[i])>=Tr
5. Off_np = {Dp ∩ Np[i]}
6. End
7. Shuffle(Np[i], off_ np, Dp, μ)
8. Loop p Є Dp
9. Np[i] = Np[i] +1
10. End
11. End

In above algorithm
Tr: value calculate by eq. 1
Np: Negative pattern
n: number of negative pattern in document
off_np: Offender value in particular negative pattern
μ: coefficient value

4. Experiment and Result

To evaluate this work Research paper and news articles dataset is use for evaluation. As research paper are of different fields so keyword set for different field of research paper has to create and update in order to get efficient results for text document clustering this can be understand as the research paper of computer science field has the keyword set { Text mining, text classification, pattern mining, pattern evolving, information filtering,……etc.}.

In similar fashion text document of news articles is also use for evaluation of this project, and here also keyword set of different news area need to create separately this can

be understand as let the news articles on religion then its keyword set like a { atheistic, God, Heaven, Jesus, marriage, bless, love dead,……etc. }. With these different dataset work is done.

Measures

To test our result we use following measures are the accuracy of the text mining approach, that is to say Precision, Recall and F-score.
Precision = true positives / (true positives+ false positives)
Recall = true positives / (true positives +false negatives)
F-score = 2 * Precision * Recall / (Precision + Recall)

In above true positive means that the submit positive document is identify as positive document and false negative means submit positive document is identify negative document and vice versa. False Positive means submit negative document is identifying as positive.

Table Represent values after experiment:-

Dataset of 100 document	Precision	Recall	F-Score
Research Paper	0.81	0.92	0.861
News Articles	0.79	0.90	0.841

From above table it is seen that accuracy for document finding by using pattern mining with the help of keywords give an effective results.

5. Conclusion

As text mining is done by different way such as term, phrase and pattern base. In this work one concept of pattern base is use with a mix of keywords of that document, this term has given an effective results that are highly dynamic, as it was acceptable for paper of different field. In this work the approach of developing the effective pattern is done both with the help of relevant or positive document or irrelevant or negative document and a set of keywords that contain text of that relevant fields. As negative document will reset the weights as well. This work accept document from any field like research paper or news articles although keywords will be different and it was shown in results that more than 80 percent of document are successfully identify.

References

[1] S. R. Sharma and S. Raman, “Phrase-Based Text Representation for Managing the Web Document,” Proc. Int’l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.
[2] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo,“Applying Data Mining Techniques for Descriptive PhraseExtraction in Digital Document Collections,” Proc. IEEE Int ’lForum on Research and Technology Advances in Digital Libraries(ADL ’98), pp. 2-11, 1998.
[3] D.D. Lewis, “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task,” Proc. 15th Ann. Int’l ACM
[4] SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’92), pp. 37-50, 1992.
[5] Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.
[6] Manning and H. Schu¨ tze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
[7] 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.
[8] Y. Li and N. Zhong, “Interpretations of Association Rules by Granular Computing,” Proc. IEEE Third Int’l Conf. Data Mining
[9] (ICDM ’03), pp. 593-596, 2003.
[10] Y. Xu and Y. Li, “Generating Concise Association Rules,” Proc.ACM 16th Conf. Information and Knowledge Management (CIKM ’07), pp. 781-790, 2007.
[11] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, “Automatic Pattern-Taxonomy Extraction for Web Mining,” Proc. IEEE/WIC/ACM Int’l
[12] S.-T. Wu, Y. Li, and Y. Xu, “Deploying Approaches for Pattern Refinement in Text Mining,” Proc. IEEE Sixth Int’l Conf. Data
[13] 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