Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

A Survey on Pattern Taxonomy Model Technique for Retrieving Pattern to Classify Text Document

Mukund N.Patil[1], Prof.Priti Subramanium[2]
[1] *C.S.E.Dept.S.S.G.B.C.O.E.T,Bhusawal*
[2] *C.S.E.Dept.S.S.G.B.C.O.E.T,Bhusawal*

Abstract— when a large number of research document are received, it is common to group them according to their similarities in research disciplines. The grouped document is then assigned to the appropriate experts for peer review. Current methods for grouping document are based on manual matching of similar research discipline areas and/or keywords. However, the exact research discipline areas of the document cannot often be accurately designated by the applicants due to their subjective views and possible misinterpretations. Therefore, rich information in the document full text can be used effectively. Text-mining methods have been proposed to solve the problem by automatically classifying text documents, mainly in English. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This work presents an innovative and effective pattern retrieving technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating retrieved patterns for finding relevant and interesting text document.

Keywords- Computer Networks, Network Security, Anomaly Detection, Intrusion Detection

I. INTRODUCTION

Text Mining 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 analyses 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 Are as 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.

II. 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. , 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 proposed in .In 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. In, hierarchical clustering 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. 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 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 and 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 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 was presented to bridge the gap between NLP and text mining, which analysed terms on the sentence and document levels. pattern based methods was introduced in to significantly improve the performance of information filtering.

modules of the paper are

a. Text Pre-processing
b. Pattern taxonomy process
c. Pattern deploying
d. Pattern evolving

Text Pre-Processing

Data pre-processing 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 paper 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}.

Pattern taxonomy process

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. 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)

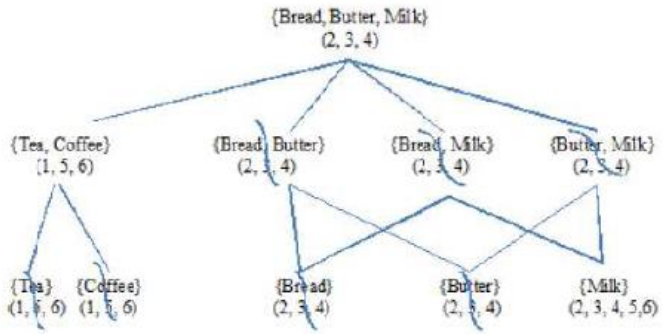
Table 1.

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

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)|tЄSp[i]}
6. d = d + p
7. end
8. Dp = Dp U d
9. End
10. Loop p ЄSp
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.

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 pЄDp
9. Np[i] = Np[i] +1
10. End
11. End

III. CONCLUSIONS

Using Pattern taxonomy model technique we retrieve the pattern for classifying or separation of text document from a large number of text . Pattern taxonomy model technique solve the problem of misinterpretations related to text mining

REFERENCES

[1] Mukund N.Patil & Prof.Priti Subramanium "Pattern Taxonomy Model Technique for Retrieving Pattern to Classify Text Document" International Journal of Application or Innovation in Engineering & Management (IJAIEM) at volume 4 & issue no 2 February 2015.

[2] 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

[3] 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.

[4] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM.

[5] SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), pp. 37-50, 1992

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

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

[8] 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.

[9] Y. Li and N. Zhong, "Interpretations of Association Rules by Granular Computing," Proc. IEEE Third Int'l Conf. Data Mining

[10] (ICDM '03), pp. 593-596, 2003.

[11] Y. Xu and Y. Li, "Generating Concise Association Rules," Proc.ACM 16th Conf. Information and Knowledge Management (CIKM '07), pp. 781-790, 2007.

[12] 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

[13] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data

[14] 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.

