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

Efficient Pattern-Based Query search in Text Documents

Rosemary Varghese[1], Kala Karun[2]
[1, 2] Computer Science and Engineering
Adi Shankara Institute of Engineering and Technology
Kalady, India

Abstract: Text mining is a technique that helps the user to find useful information from a large amount of digital text document. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. Most existing text mining methods adopted term-based approaches, but they all suffer from the problem of polysemy and synonymy. Text mining is a technique that helps the user to find useful information from a large amount of digital text document. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. Most existing text mining methods adopted term-based approaches, but they all suffer from the problem of polysemy and synonymy. The effective usage and updating of discovered patterns is still an open research issue. Pattern deploying and pattern evolving method has also been proposed in order to refine the patterns that help in improving the effectiveness of pattern discovery. There are two phases that we need to consider when we use pattern-based models in text mining: one is how to discover useful patterns from digital documents, and the other is how to utilize these mined patterns to improve the system's performance. The new approach use pattern (or phrase)-based approaches which perform better in comparison studies than other term-based methods. It uses a pattern taxonomy model. In pattern taxonomy model, given documents are separated into different paragraphs.

Keywords— Pattern based approach, pattern Taxonomy model, pattern mining, Information retrieval

I. INTRODUCTION

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. 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 Research Work use pattern (or phrase)-based approaches which perform better in comparison studies than other term-based methods.

The advantages of termbased methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, termbased 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. The semantic meaning of many discovered terms is uncertain for answering what users want. The disadvantage of phrase-based approaches are 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence, and 3) there are large numbers of redundant and noisy phrases among them.

This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving. It divides the documents in to positive and negative documents based on the user search query. It identify the pattern from the documents and the result is displayed to the user

II. STATE OF ART

1) Rough set decision rule-Based topic filtering

Xujuan Zhou, Yuefeng Li, Peter Bruza, Yue Xu and Raymond Lau proposed the algorithm [4]. Rough set decision rule-Based topic filtering. The decision rules for the partitioning of the incoming document stream into the positive, boundary and negative regions have been developed in this model. The proposed two-stage IF system uses the strategies used in both batch filtering and routing filtering. In the first stage, a topic filtering method based on the Rough Set decision rules is used to develop an optimal threshold. All the unlikely relevant documents are filtered out. The remaining documents of the incoming stream will pass into the second stage. The pattern mining method at the second stage will work on only a relatively small amount of documents. The remaining documents are potentially with a higher relevance at the second stage. Thus, better ranking accuracy will yield in the routing filtering process

2) Concept based extractor algorithm

Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau proposed this method [5]. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled term either word or phrase is considered as concept. The extracted top concepts are used to build standard normalized feature vectors using the standard vector space model (VSM) for the purpose of text categorization cost.

3) Pattern deploying method (PDM) & Pattern deploying with relevance feedback (PDR)

S.-T. Wu, Y. Li, and Y. Xu, proposed this algorithm [6].PDR is a deploying method with the use of relevance function. This approach utilizes a probabilistic method to estimate the term weight. In order to deploy discovered patterns and acquire deployed support of each term in these patterns, it uses a pattern composition operation to join two patterns. This paper uses a pattern deploying strategies using pattern deploying method and pattern deploying with relevance feedback.

4) Class sequential rule(CSR)

N. Jindal and B. Liu, proposed this method [7]. This paper studies the problem of identifying comparative sentences in text documents. Extracting comparative sentences from text is useful for many applications. For example, in the business environment, whenever a new product comes into market, the product manufacturer wants to know consumer opinions on the product, and how the product compares with those of its competitors. Much of such information is now readily available on the Web in the form of customer reviews, forum discussions, blogs, etc. Extracting such information can significantly help businesses in

their marketing and product benchmarking efforts. This paper focus on product comparisons those are not only useful for product manufacturers, but also to potential customers as they enable customers to make better purchasing decisions. This paper classifies comparative sentences into different categories based on existing linguistic research. This paper focus on the study the problem of identifying comparative sentences in text, a categorization of comparative sentences into different types, effective approach to solve the problem based on class sequential rules and the machine learning technology

5) Semantic relation identification &Topic specificity.

Y. Li and N. Zhong proposed this model [8].In the web information gathering, user profiles were used to understand the semantic meanings of queries and capture user Information needs. User profiles are used for user modeling and personalization. It is used to reflect the interests of user. The user profiles are categorized into two diagrams: the data diagram and which are acquired by analyzing a database or a set of transaction whereas the information diagram user profiles acquired by using manually such as questionnaires and interviews or automatic techniques such as information retrieval and machine learning. User profiles are categorized into three groups: interviewing, semi-interviewing, and non-interviewing

6) SPmining

S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen proposed this method [10]. It present a new pattern-based model PTM (Pattern Taxonomy Model) for the representation of text documents. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. The "pattern" used as a word or phrase in this paper is extracted from the text documents. SPMining adopts the concept of projected database method for extracting frequent sequential patterns from a document. The main difference between SPMining and others which adopt the same concept is that SPMining deals with several sequences at a time, whereas others only handle one sequence at a time

III. METHODOLOGY

The method presents a pattern-based text-mining Approach. The method is efficient and effective for identifying patterns with English language texts. The method also includes a pattern taxonomy model to discover patterns and pattern deploying methods to update discovered patterns based on their frequency

The basic modules of this project includes:

- Preprocessing
- Pattern Taxonomy model
- Pattern Deploying based on the supports

Preprocessing

All words passes to preprocessing level. Irrelevant terms are eliminated there. This process is also called as *tokenization* process. It consists of two kinds operations such as stop list removal, stem word removal.

A). Stop List Removal*:* Stop words are words which are filtered out prior to, or after, processing of natural language data. They typically comprise prepositions, articles, and so on. There is no specific list of stop words for all applications and these stop words are controlled by the human but not automated. It saves the system resources. Stop word has list of words. That are deemed or irrelevant and then it is removing .It consists of articles (a, an, the), preposition (for, in, at, etc.), and so on.

B). Stem Word Removal*:* Stemming is the process for
reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms. In this preprocess the text documents have to be processed using the Porter stemmer. It

removes the Suffix‟s of the words these words are useful in the text mining for clustering the text documents in the text mining process we collects the documents and each documents are composed into the set of terms or words the words having stem have a same meaning in stem process the suffixes of the words, singular and plural words are considered into a one single word for meaning full text clustering process.

Pattern taxonomy model

The pattern-based model PTM (Pattern Taxonomy Model) for presenting the text data. It is a structure like prune which has frequently occurring data, as a root node and others its nodes as a subset of it in text documents. The two Basics approach regarding performance in phrases is its low frequency of occurrence and false terms. A phrase with maximum value of the support is a general term which occurs frequently and as the value of minimum support minimizes more the unidentified phrases are searched, to avoid this and to satisfy what the user actually want we have pattern taxonomy model. Pattern taxonomy is a tree like structure that shows the relation between patterns discovered from a text data. It is „IS‟ a relationship with most relevant patterns and subsequence.

For example, pattern. < P; Q > is a sub-sequence of pattern < P; Q; R > and pattern is a sub-sequence of pattern. Thus the root of the tree at the bottom level is < P; Q; R >represents frequent patterns (i.e. More sequential patterns). Once the tree is structured, we are able to find links between different phrases. Thus, in this we shrubs the meaningless pattern, i.e. it is obvious that the pattern < P; Q >occurs in every paragraph of < P; Q; R >. Hence it is considered as meaningless pattern. As in super sequence pattern more subsequence pattern can occur. The diagram below shows that the pattern occurs frequently, i.e. < P; Q >, < Q; R >, < P; R >, < P >, < R >, Q > in P, Q, R we can consider this pattern as less frequently or not useful patterns The fig 1 shows the pattern taxonomy Model The sequential patterns that we get after pruning are composed using pattern deploying method

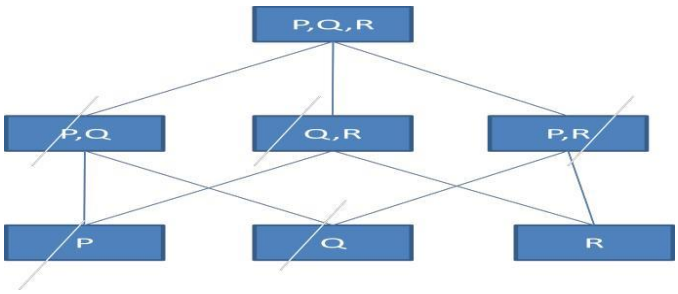

Fig 1: Pattern Taxonomy Model

Pattern deploying method

In order to use the semantic information in the pattern taxonomy to improve the performance of closed patterns in text mining, we need to interpret discovered patterns by summarizing them as d-patterns in order to accurately evaluate term weights (supports). The rational behind this motivation is that d-patterns include more semantic meaning than terms that are selected based on a term-based technique (e.g., tf*idf). As a result, a term with a higher tf*idf value could be meaningless if it has not cited by some d-patterns (some important parts in documents). The evaluation of term weights (supports) is different to the normal term-based approaches. In the term-based approaches, the evaluation of term weights are based on the distribution of terms in documents. In this research, terms are weighted according to their appearances in discovered closed patterns

Pattern deploying methods are used to overcome the low-frequency problem of specific long patterns. Deploying patterns through the use of a pattern composition operator, the goal of reserving the significant information embedded in specific patterns can be achieved. The patterns support obtained in the pattern discovery phase is taken into account when we deploy patterns to a common hypothesis space. By using SPMining algorithm we can acquire a set of frequent sequential patterns

D-pattern mining algorithm is used to discover the Dpatterns from the set of documents. The efficiency of the pattern taxonomy mining is improved by proposing an SP mining algorithm to find all the closed sequential patterns, which is used as the well-known appropriate property in order to reduce the searching space. The algorithm describes the training process of finding the set of d-patterns. For every positive document, the SP Mining algorithm is first called giving rise to a set of closed sequential patterns. The main focus is the deploying process, which consists of the d--pattern discovery and term support evaluation. All discovered patterns in a positive document are composed into a d-pattern giving rise to a set of d-patterns .Thereafter, term supports are calculated based on the normal forms for all terms in d-patterns.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining_IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING", VOL. 24, NO. 1, JANUARY 2012
[2]Sheng-Tang Wu, " Knowledge Discovery Using Pattern Taxonomy Model in Text Mining 'December 2007
[3] Kavitha Murugeshan, Neeraj RK,"Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE), May 2013
[4] 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.
[5] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Text Categorization," Proc. 13th Int"l Conf. Knowledge Discovery and Data Mining (KDD "07), pp. 629-637, 2007.
[6] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int"l Conf. Data Mining (ICDM "06), pp. 1157-1161, 2006.
[7]N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents", Proc. 29thAnn. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006
[8]Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.
[9] "Effective Pattern Deploying Approach in Pattern Taxonomy Model for Text Mining, International Conference on Recent Trends in engineering & Technology" - 2013(ICRTET'2013)
[10]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 Conf. Web Intelligence (WI "04), pp. 242-248, 2004.
[11] Bhushan Inje , Ujawla Patil,IEEE Students" Conference on Electrical, Electronics and Computer Science Operational Pattern Revealing Technique in Text Mining"- 2014
[12] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int"l Joint Conf. Artificial Intelligence (IJCAI "03), pp. 587-594, 2003.
[13] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task, "Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), Pp37-50, 1992.
[14] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization, "Proc. Workshop Speech and Natural Language, pp. 212-217, 1992
[15] K. Aas and L. Eikvil, _Text Categorizations:" A Survey, Technical Report Raport" NR 941, Norwegian Computing Center, 1999.
[16] International Journal of Advanced Computer Research "Text mining: A Brief survey "December-2012