International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Discovering Efficient Patterns for Text Mining Approach: A Survey

Priyanka R. Magar[1], C. S. Biradar[2]

[1]SRTM University, Department of CSE, M. S. Bidve Engineering College, Latur, Maharashtra, India

[2]Visvesvaraya Technological University, Department of CSE, B.E.C. College of Engineering, Bagalkot, Karnataka, India

Abstract: ***Text mining is technique which extracts interesting knowledge in various text documents. Approximately 90% world's data is held in unstructured formats. In order to get useful information from this data there is great need for mining techniques. There are many techniques for mining the useful patterns from text documents, such as frequent itemset mining, closed pattern mining etc. Most of the existing text mining techniques are built on term based approach, but it faces problems of synonymy and polysemy. Pattern based approach also not produce much efficient result. In this survey paper, we focused on developing efficient mining algorithm for discovering patterns from text collections and search for useful and interesting patterns by using pattern deploying and pattern evolving.***

Keywords: Sequential pattern mining, Text mining, Pattern evolving, Pattern deploying.

1. Introduction

Knowledge discovery has become preeminent phenomenon in recent years due to the rapid increase in digital data. It consists of numerous methodologies, used for extorting useful knowledge from unstructured and structured data. Many applications, such as market analysis and business management, can profit by the use of the information and knowledge extracted from a large amount of data. Data mining is therefore a vital step in the process of knowledge discovery in databases. In the past years, most important variety of data mining techniques are presented in order to carry out different knowledge tasks. These techniques embrace association rules mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are anticipated for the purpose of developing efficacious mining algorithms to find particular patterns within a reasonable and acceptable time frame [2]. With a large number of patterns produced by using the data mining approaches, how to effectively exploit these patterns is still an open research issue.

Text mining is the discovery of impressive knowledge in text documents. It's difficult issue to find out proper data in text documents to help users to seek out what they want. It is a demanding work to use those patterns and also bring them up to date. Earlier term based methods are provided by Information Retrieval (IR) techniques. The term based methods are organized into probability models [3], rough set models [4] and SVM based models [5]. All term based methods suffer from troubles such as polysemy and synonymy. When a word has a variety of meanings, it is known as polysemy. When various words have the equivalent meaning, it is called synonymy. Thus the semantic meaning of various discovered terms are unpredictable for answering what users want.

Because of this reason, many years people believed that phrase based techniques are better than that of term based technique. However, the experiments in the field of data mining [6], [7] have not been proved. The possible reason include the phrases have less properties pertaining to statistics when compared with terms; frequency of occurrence is low; noisy and redundant phrases are more. Though there are some disadvantages, the sequential patterns turn out to be capable alternatives to phrases. Pattern Taxonomy Model (PTM) is used to overcome the drawbacks of phrase based approaches. Pattern based approaches became alternatives but much improvements are not made to make them more effective for text mining. With regard to effectiveness there are 2 issues. They are misinterpretation and low frequency. When patterns are not as a large amount of frequent, they can't be used for conclusion. When the terms or patterns are misinterpreted, the end product will not be reliable.

In order to solve the problem mentioned above, this survey paper presents a novel pattern discovery technique proposed by Zhong et al. [2] which first measures the specificities of discovered patterns and then evaluates weights of terms according to distribution of terms. Thus it avoids misinterpretation problem. It also assumes negative training examples to find noisy patterns and tries to avoid their influence for the low frequency problem.

2. Related Work

2.1 Text Mining

The purpose of text mining is to process unstructured data, extract meaningful information in text collection. Information can be extracted to derive summaries for the words contained in the documents. Hence, we can analyze documents and determine similarities between them. Text mining also referred as text data mining, roughly equivalent to text analytics, it refers to procedure of deriving high quality of information from text and high quality of information is derived through devising of patterns. Text analysis involves information retrieval, lexical analysis, word frequency distributions, pattern recognition, information extraction and

data mining techniques. In [2] adopted several approaches for text mining which showed in Table1.

Table 1: Comparison of Text Mining Methods

Sr. No.	*Type Of Method*	*Algorithm*	*Feature Selection*
1	Term Based Method	TFIDF	Term
2	Phrase Based Method	SP Mining	Phrase
3	Pattern Based Method	PDM	Pattern

2.2 Frequent and Closed Patterns

In terms of pattern discovery, data mining techniques can be used for pattern discovery. Frequent patterns are one that occurs in at least a user specific percentage of database, that percent is called support. Consider d as document, PS(d) is set of paragraphs in that document, X is term set. $\lceil X \rceil$ is used to denote the covering set of X for d, which inserts all paragraphs dp Є PS(d) such that $X \subseteq$ dp i.e. X={dp | dp Є PS(d), X ⊆ dp}. Its absolute support is the number of occurrences of X in PS(d), that is sup_a(X) = $|\lceil X \rceil|$. Its relative support is the fraction of the paragraphs that contain the pattern which is, sup_r(X) = $\frac{X}{PS(d)}$. A term set X is referred as frequent pattern if its sup_a (or sup_r) ≥ min_sup, a minimum support. The closure of X is defined as, Cls(X) = termset($\lceil X \rceil$). A pattern is called closed pattern if X= Cls (X).

2.3 Pattern Taxonomy Model (PTM)

In PTM, documents split into set of paragraphs and each paragraph consists of set of words. At this stage, apply the data mining technique to find frequent patterns and generate pattern taxonomies. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. Also, the weight of the term which is occurring in extracted pattern is calculated. The PTM discovered closed sequential patterns and pruned nonclosed patterns in the text documents. For the duration of the pruning phase, non-meaning and redundant patterns are eliminated by applying pruning scheme and we get meaningful patterns. The overview of discovering patterns are mentioned in Figure1, where text document is given input to SP Mining and patterns are generated.

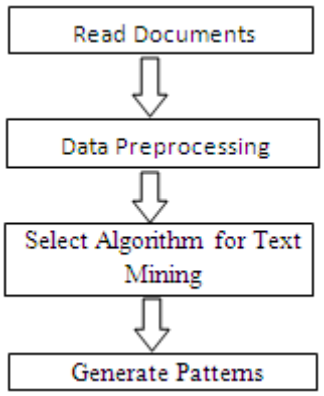

Figure 1: Flow Diagram of Pattern Discovery

2.4 Pattern Deploying Method (PDM)

The main focus of the paper is deploying process which consists of- d pattern discovery and term support evaluation. The worth of patterns can be projected by assigning an evaluated value based on one existing function. In [8] pattern deploying methods are anticipated for the use of knowledge discovery. All discovered patterns are not interesting because a number of noise patterns are also extracted from the training dataset. Information from the negative sample is not demoralized during that concept learning. The negative document also contains useful information to identify ambiguous pattern in the concept. It is easier to locate the associated document if the same patterns appear in the positive document. But if the analogous pattern appears in the negative document it will be complicated. To enlarge the effectiveness it is indispensable for a system to exploit ambiguous pattern from the negative examples in order to decrease their influence.

2.5 Pattern Evolving Method

Patten evolution is used to identify the noisy patterns in documents and update d-pattern by shuffling. This technique helps to reduce the effects of noisy patterns because of the low frequency problem. This method is called inner pattern evolution because it only changes a pattern's term supports within pattern. A threshold is used to categorize documents into relevant or irrelevant categories. In order to diminish the noise, d-patterns are tracked and find out which pattern give rise to such an error [11]. These patterns are offenders. There are two types of offenders complete conflict offenders and partial conflict offenders, the idea of updating patterns are explained here: Firstly, complete conflict offenders are deleted from discovered d-patterns then for the partial conflict offenders reshuffling of their term support is carried out in order to reduce the effects of noisy documents. This algorithm gives better result and efficient updating of discovered patterns which are extracted from text documents.

3. Conclusion

In this paper we have presented review of pattern discovery technique for text mining approach. There are various data mining techniques proposed in last year's but, the updating of discovered pattern effectively was difficult with those techniques because the long pattern with high specificity lacks in support. Insufficient use of patterns that are extracted also causes performance degradation. The main issue regarding the pattern based approach is low frequency and misinterpretation. But, this paper presents pattern deploying and pattern evolution method to solve these issues.

References

[1] K. Aas and L. Eikvil, "Text Categorization: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.
[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining" VOL. 24, NO. 1, JANUARY 2012.

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
[4] Y. Li, C. Zhang and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.
[5] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.
[6] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
[7] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
[8] 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
[9] 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.
[10] 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.
[11] S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" In iiWAS'05, pages 1013–1022, 2005.

Author Profile

P.R. Magar received the B.E. degree in Computer Science and Engineering from M.S. Bidve Engineering College in 2012. Now, she is pursuing Master's in Engineering (Computer Science and Engineering) from M.S. Bidve Engineering college, Latur , SRTM University Nanded, Maharashtra.

C.S. Biradar received the B.E. and M.Tech. degrees in Computer Science & Engineering from B. L. D. E. A College of Engineering, Bijapur in 2011 and from BEC College of Engineering, Bagalkot in 2013, respectively. He is now with M.S. Bidve Engineering College, Latur (M.S.) as Assistant Professor since 2013.