IJCSN International Journal of Computer Science and Network, Volume 4, Issue 2, April 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

Pattern Mining Technique for Text Mining

[1] Pragati Dubey, [2] Prashant Dahiwale

[1] Department of Computer Science & Engineering, RTMNU, RGCER
Nagpur, Maharashtra, India

[2] Department of Computer Science & Engineering, RTMNU, RGCER
Nagpur, Maharashtra, India

Abstract - Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. 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 (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. In This approach we used an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

Keywords - *Text Mining, Pattern Mining, Pattern Taxonomy, Pattern Evolution.*

1. Introduction

Due to the rise of knowledge created obtainable in recent years, information discovery and data processing have attracted a good deal of attention with associate close need for turning such knowledge into helpful data and knowledge. Several applications, like market research and business management, will profit by the employment of the information and information extracted from an outsized quantity of data. Information discovery will be viewed because the method of nontrivial extraction of data from massive databases, information that's implicitly conferred within the knowledge, previously unknown and probably helpful for users. Data mining is so a vital step within the method of knowledge discovery in databases.

1.1 Text Mining

Text mining, typically alternately observed as text data processing, roughly akin to text analytics, refers to the method of explanation high-quality data from text. High-quality data is usually derived through the making of patterns and trends through suggests that like applied Statistical pattern learning. Text mining typically involves the method of structuring the input text explanation patterns among the structured information, and at last analysis and interpretation of the output. 'High quality' in text mining typically refers to some combination of connexion, novelty, and interest. Typical text mining tasks; text categorization, text cluster, sentiment analysis, production of granular taxonomies, concept/entity extraction, document account, and entity relation modelling (i.e., learning relations between named entities).

Text analysis involves data retrieval, lexical analysis to review word frequency distributions, tagging/annotation, pattern recognition, data extraction, data processing techniques together with link and association analysis, and prophetic analytics. The overarching goal is, basically, to show text into information for analysis, via application of linguistic communication process and analytical strategies. A typical application is to scan a group of documents written in an exceedingly linguistic communication and either model the document set for prophetic classification functions or populate a information or search index with the knowledge extracted.

Text mining is that the discovery of fascinating information in text documents. it's a difficult issue to search out correct knowledge (or features) in text documents to assist users to find what they require. Within the first, info Retrieval (IR) provided several term-based strategies to resolve this challenge, like Rocchio and probabilistic models [4], rough set models [2], BM25 and support vector machine (SVM) [4] based mostly filtering models. the benefits of term based methods embrace economical process performance as well as mature theories for term weight, which have emerged

over the last few decades from the IR and machine learning communities. However, term based methods suffer from the issues of equivocalness and synonymy, wherever equivocalness suggests that a word has multiple meanings, and semantic relation is multiple words having the same which means. The linguistics which means of the many discovered terms is unsure for respondent what users wish. Over the years, folks have typically command the hypothesis that phrase-based approaches may perform higher than the term based ones, as phrases could carry additional "semantics" like information. This hypothesis has not fared too well within the history of IR . though phrases area unit less ambiguous and additional discriminative than individual terms, the probably reasons for the discouraging performance include: 1) phrases have inferior applied mathematics properties to terms, 2) they have low frequency of incidence, and 3) there are a unit giant numbers of redundant and creaking phrases among them .

1.2 Pattern Mining

"Pattern mining" is Information Hiding. In data mining method that involves finding existing patterns in data. During this context patterns typically means that association rules. an oversized system needs decomposition. a way to decompose a system is to phase it into collaborating objects. In massive systems a first-cut rough model would possibly turn out lots of or thousands of potential objects. Further refactoring generally results in object groupings that offer connected kinds of services. once these teams square measure properly segmental, and their interfaces consolidated, the result's a superimposed design.

In the past decade, a major variety of knowledge mining techniques are given so as to perform different information tasks. These techniques embody association rule mining, frequent item mining, sequential pattern mining, and closed pattern mining. With an outsized variety of patterns generated by victimization information mining approaches, a way to effectively use and update these patterns continues to be AN open analysis issue. During this paper, focus on the event of a information discovery model to effectively utilize and modernize the discovered patterns and apply it to the sphere of text mining.

2. Related Work

In [3], data processing techniques are used for text analysis by extracting coincidental terms as descriptive phrases from text collections. Conversely, the effectiveness of the text mining systems exploitation phrases as text representation showed no vital improvement. The likely reason was that a phrase-based technique had "lower consistency of assignment and lesser document frequency for terms" as mentioned in [4]. Term-based metaphysics mining ways additionally provided some thoughts for text representations. as an example, hierarchical clustering [8], [2] was wont to verify synonymy and semantic relation relations among keywords. Also, the pattern progress technique was introduced in [5] so as to improve the act of term-based metaphysics mining.

Pattern mining has been extensively considered in knowledge mining communities for several years. a spread of economical algorithms like apriori-like algorithms [2], prefix span fp-tree,spade are projected. These analysis works have chiefly centered on developing economical mining algorithms for discovering patterns from an outsized knowledge assortment. However, checking out helpful and attention-grabbing patterns and rules was still associate in open downside [6], within the field of text mining; pattern mining techniques are often wont to notice various text patterns, like ordered patterns, frequent item sets, coincidental terms and several grams, for building up an illustration with these new varieties of options.

Nevertheless, the difficult issue is the way to effectively deal with the massive quantity of discovered patterns. For the difficult issue, closed ordered patterns have been used for text mining in [5] that projected that the concept of closed patterns in text mining was helpful and had the potential for raising the performance of text mining. Pattern taxonomy model was additionally developed to improve the efficiency by effectively exploitation closed patterns in text mining. Moreover, a two-stage model used each term-based strategies and pattern based methods was introduced in to considerably improve the performance of data filtering.

3. System Model

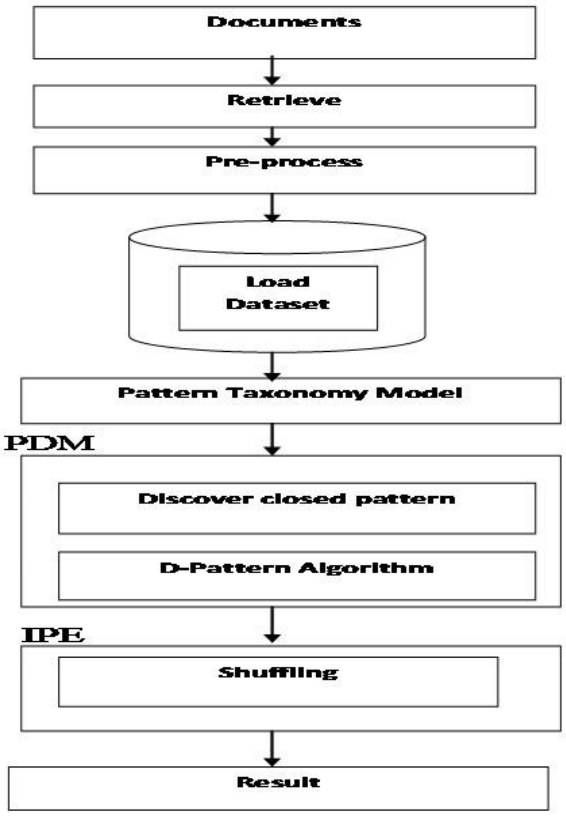

Fig.1 Overall System Model

4. Proposed System

4.1 Pattern Taxonomy Model(PTM)

Two main stages are considered in PTM. The first stage is how to extract useful phrases from text documents, which will be discussed in this chapter. The second stage is then to use theses discovered to improve the effectiveness of a knowledge discovery system and will be presented.

In PTM, we split a text into set of paragraphs and treat each paragraph as an individual transaction, which consists of a set of words. At the subsequent phase, apply the data mining method to find frequent pattern from these transaction and generate pattern taxonomies. During the pruning phase, non-meaning and redundant pattern are eliminated by applying a proposed pruning scheme.

4.2 Pattern Pruning

For all existing algorithm used for finding all sequential pattern from dataset, the problem encountered is large amount of patterns generated, most of which are considered as non-meaningful pattern and need to be eliminated. A proper pruning scheme can be used for addressing this issue by removing redundant patterns, leading to not only reducing the dimensionality but also decreasing the effect from noise patterns. In this research work, defined closed patterns as meaningful pattern since most of the sub-sequence pattern of closed pattern have the same frequency, which means they always occur together in the document. For example fig2 ,pattern <t1, t2>and<t1, t3>appear two times in a document as there pattern<t1,t2,t3>has a frequency of two.SPM stands for sequential pattern and we defined sequential closed pattern mining as SCPM.The algorithm and setting minimum support to be 0.6, a list of the entire closed or non-closed sequential pattern can be returned and their results are shown in table 1.

Table 1 : Patterns

Pattern	Non-closed Pattern	Closed Pattern
Dp1	<t2><t4>	<t1><t3><t5>
Dp2	<t1,t2><t1,t3>	<t2,t3><t2,t4> <t5,t3>
Dp3	none	<t1,t2,t3>

4.3 Using Discovered Pattern

The next issue is how to use these discovered patterns. There are various ways to utilize discovered pattern by using a weighting function to assign a value for each pattern according to its frequency. One strategy has been implemented and evaluated, which proposed a pattern mining method that treated each found sequential pattern as a whole item breaking it into a set of individual terms, and its result found that using confidence as the pattern measure outperformed the use of support.

4.4 Pattern Deploying Method

In this, propose two novel approaches with the attempt of addressing the drawbacks which is caused by the inadequate use of discovered patterns. The properties of pattern(support and confidence)used by data mining based method in the phase of pattern

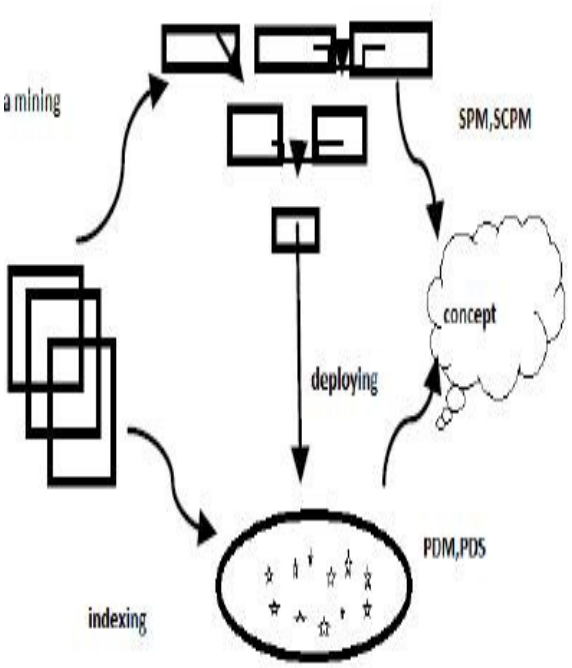

Fig 2. Pattern Deploying approach

discovery are not suitable to be adopted in the phase of using discovered pattern .therefore in this ,reevaluate the property by deploying them into a common hypothesis space based on their correlation to the pattern taxonomies. A fundamental mechanism, pattern deploying method, is firstly introduced to implement patterns deploying and followed by the method of pattern deploying based on support (PDS).

To use these patterns, two inevitable issues arise:

1) How to emphasis the significance of specific patterns and avoids the low-frequency problem.
2) How to eliminate the interference by the general patterns, which are usually with high frequency.

4.5 Pattern Evolving

In this section, we tend to discuss a way to reshuffle supports of terms at intervals traditional styles of d-patterns supported negative documents within the coaching set. The techniques are going to be helpful to reduce the aspect effects of streaky patterns thanks to the low-frequency drawback. This method is named inner pattern evolution here, as a result of it solely changes a pattern's term supports at intervals the pattern. A threshold is typically accustomed classify documents into relevant or orthogonal classes. Exploitation the d-patterns, the threshold may be outlined.

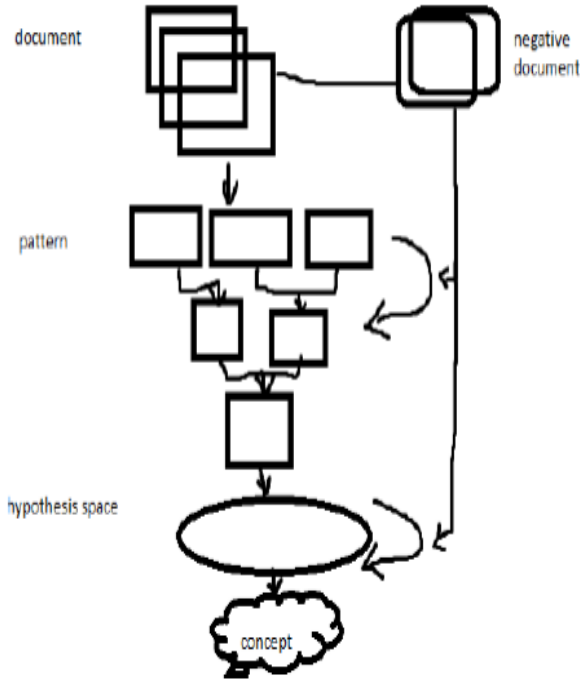

Fig.3 Pattern Evolving Approach

5 Algorithm and Discussion

5.1 Procedure

1) System starts from one of the topic and retrieve from dataset with regard to training set, such as file list and the number of documents.

2) Each document is preprocessed with word stemming and stopword removal and transformed into a set of transaction based on its nature of document structure.

3) Systems select one of pattern discovery algorithm to exact pattern. Discovered pattern are deployed into a hypothesis space using one of the proposed deploying methods. If required, the pattern evolving process is used to refine patterns.

4) A concept represent the context of the topic is eventually generated. Each document in the test set is assessed by the document evaluation method and the experimental result .system ends

for this topic and repeats the above steps for the next topic if required.

5.1 Pattern Discovery

The SCPM method is chosen as a mining mechanism in order to find frequent sequential closed patterns from transactions. Each document now is represented by pattern taxonomies which consist of discovered patterns.

5.2 Pattern Deployment

Pattern evolving method DPE and IPE undertake different processes in this step. For DPE, the deployment of pattern is processed as usual and deployed patterns are generated and passed to the subsequent step. However for IPE, there is no need for patterns to be deployed before they are evolved. In terms of pattern deploying, either PDM or PDS can be selected to perform the task.

5.3 Pattern Evolution

There are two approaches for pattern evolution, DPE and IPE.Both approaches need the information from the negative documents ("nds"). The DPE method evolves patterns based on the deployed patterns which are viewed as term level evolution, whereas the IPE method processes the task directly on the non deployed patterns, the results from the step of pattern discovery, which is referred to as pattern level evolution.

In relationship of the development as inner evolving is applied and therefore the abovementioned price of Ratio. As we are able to see that the degree of improvement is in direct proportion to the score of magnitude relation. Meaning the additional qualified negative documents square measure detected for idea revision, the additional improvements are capable to attain. In other words, the expected result will be achieved by mistreatment the proposed approach.

6 Conclusion

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective.

The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem), not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. An effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

References

[1] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.

[2] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[3] "Knowledge Discovery in Text Mining Technique Using Association Rules Extraction" Bhujade .V,janwe CICN, 2011

[4] 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.

[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C.Watkins, "Text Classification Using String Kernels, "J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[6] F. Sebastiani, "Machine Learning in Automated Text Categorization, "ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[7] M. Seno and G. Karypis, "Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint," Proc. IEEE Second Int'l Conf. Data Mining (ICDM '02), pp. 418-425, 2002.

[8] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf.Data Mining (ICDM '06), pp. 1043-1048, 2006.

[9] Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.

[10] 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

[11] Y.Yan and Y.Li "Generating concise association rules" Proc.ACM 16th Conf. Information and Knowledge Management (CIKM '07),pp. 781-790, 2007pp.

[12] 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

Authors

Pragati Dubey Completed MCA from Ramdeo Baba College of Engineering & Management, Nagpur, Maharashtra, India. Pursuing MTech 4th sem from Rajiv Gandhi College of Engineering & Research, Nagpur, Maharashtra, India. Her research interests are in the area of the Text mining, Database.

Prashant Dahiwale Pursuing PhD, Pass out MTech from VNIT, Nagpur, Maharashtra, India. His research interests are in the area of the Web Crawler.