Efficient Updating of Discovered Patterns for Text Mining: A Survey

Anisha Radhakrishnan
Post Graduate Student
Karunya university
Coimbatore, India

Mathew Kurian
Assistant Professor
Karunya University
Coimbatore, India

ABSTRACT

Text mining is the techniques of retrieving interesting information from the text document. Through the devising of patterns, we can retrieve high-quality information. There are many techniques for mining the useful patterns from the text document. Researchers are still going in efficient updating of discovered pattern. Polysemy and synonymy are the problem faced in term based approach. Phrase based approach also did not provide the efficient results. This paper presents an outline of effectiveness of using and updating patterns for finding interesting and relevant information from the text document by using two methods pattern evolving and deploying.

Keywords

Text mining, text classification, pattern mining, pattern deploying, pattern evolving

1. INTRODUCTION

Knowledge discovery and data mining consist of several methodologies, used for extracting useful knowledge from data. The quick growth of online information due to the Internet and the extensive use of databases have created a vast need for KDD methodologies. There are several challenges in retrieving knowledge from data draws upon research in databases, pattern recognition, machine learning, statistics, data visualization, optimization, and high-performance computing [16]. It helps to provide advanced business intelligence and web discovery solutions. Discovered knowledge is the output from the system that extracts pattern from the set of fact from the database.

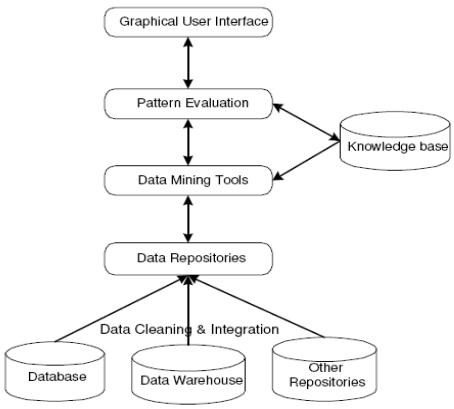

Fig 1: Knowledge Discovery in Database

Data mining is the method of pattern discovery in a data set. Knowledge discovery method is often makes it possible to use domain knowledge to guide and control the process and help to evaluate the patterns [6].Many types of data mining techniques are used association rule mining, sequential pattern mining and closed sequential pattern etc.

2. ASSOCIATION RULE

Business organizations consist of the huge amount of data for their day to day operations. Association rule are interesting pattern and it is discovered from the data set. It helps to find out frequently co-occurring elements by analyzing the data. The two common measures of interestingness are *support* and *confidence*. These measures play a vital role in the business field because low support and confidence will not be profitable. The estimation of utility of function in order to define the usefulness of mined pattern is rule support. It is calculated by the percentage of the task relevant data transaction for which the pattern is recognized as true. Certainty and validity of mined pattern is considered as confidence [6].

Support: The rules with support sup in T, the transaction data set if sup% of transactions is X ∪ Y.

Confidence: The rule with confidence in T if conf% of transaction that is X also contain Y.

- sup = Pr(X ∪ Y)
- conf = Pr(Y | X)

2.1 Association Rule Mining

Data mining plays a vital role in the field of industry in improving the marketing. E.g. is barcode strategy, retail organization. The most significant problem in data mining is mining associative rule [5].When both minimum support threshold and minimum confidence threshold are met, the association rules are considered to be interesting and useful. Associative relationship between objects is discovered in associative rule. There are two key issues in mining association rule from a large database. First issue is about generation of frequent itemset, the objective is to find out all frequent itemset that satisfy minimum support. Second one is the rule generation, helps in extracting the high confidence rule from the frequent itemset. It can be the decomposed threshold for different level of abstraction. Candidate set generation is expensive [3].

2.2 Sequential Pattern Mining

Sequential patterns are the sequences whose support exceeds the minimal support which is defined by the user. Apriori

property in association rule mining was the earlier algorithm used for sequential pattern mining. The order of the transaction that is occurring frequently in a dataset is not considered [1]. An association rule mines the intra transaction pattern and sequential patterns are to mine inter-transaction pattern. ApproxMAP (APPROXimate Multiple Alignment Pattern Mining) are developed to find approximate sequential patterns shared by many sequence, it covers many short patterns. Apriori property states that any sub-patterns of frequent patterns should be frequent. Based on this, series of Apriori- like algorithm is proposed [9]. GSP is an extension of Apriori model, uses "Generating-Pruning" method. PSP (Prefix Span) is another method based on "Generating-Pruning" principle. Candidates and frequent sequences are managed in more efficient structure in PSP than GSP [2].Another algorithm used for extracting sequential patterns is SPADE [10]. The main idea behind this method is a clustering of the frequent sequences based on their enumeration of the candidate sequences and common prefixes.

2.3 Closed Sequential Pattern

Sequential pattern included in no other sequential pattern that has the same support exactly is closed sequential pattern. ColSpan is the first algorithm designed to extract the closed sequential patterns [18]. Non-closed sequential patterns are detected avoiding the large number of recursive calls. It is based upon the detection that frequent sequences of length two such a way "A always occurs before/after B." BIDE (BI-Directional Extension) extends the previous one. It grows the prefix patterns as well as checks the closure property. It proposes BackScan pruning method to prune the search space more deeply. This method avoids the extending of sequence by detecting the extension that is already included in sequence.

2.4 Frequent Itemset

Frequently occurring subsets in a sequence of a set is frequent itemset. Many algorithms were proposed for extracting the frequent itemset. Apriori is the widest known algorithm. Later many Apriori modifications were proposed, DHP (Direct Hashing Pruning), DIC (Direct Itemset Counting), Sampling and Partition. The number of candidate itemset can be reduced by DHP [7]. Later DIC algorithm was proposed by Brin et al. This method helped in reducing the number of passes in database. The data base is divided into a period of specific size. The sampling algorithm helps to do two scan through the database by choosing the random sample. The random samples are picked from the database. All relatively frequent patterns are found out from the sample and result is verified with the database. The partitioning algorithm is completely different from other approaches. By using the vertical layout the database is stored in the main memory. The two covers of two subsets are intersected, and the support of itemset is computed [13]. When covers of all items are stored, it means that complete database is stored in the main memory but for large database this is impossible. So the database is partitioned to several disjoint parts. All relatively frequent parts for all itemset are generated by the algorithm, and it is merged. This gives the superset of all frequent items [8].

3. TEXT MINING

Text mining is the technique of retrieving information from the text document. It helps in finding the interesting patterns from the large database. Text mining can also be called as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery. Since most information is stored as text documents, text mining is supposed to have a high viable potential value. Both structured and unstructured data sets can be used in text mining for retrieving the interesting knowledge.

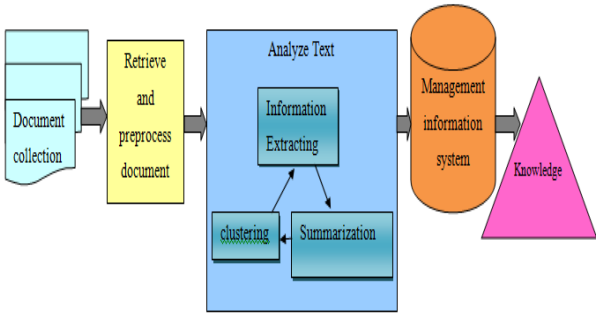

Fig 2: Knowledge Discovery in Database

This figure shows the text mining applications. Text mining is a different area from the web search. In text mining the user looks for the unstructured data while in search already known information are searched. Text classification is the tasks where the documents are assigned to predefined categories. Many text classification methods have been developed. SVM is one important machine learning method for text classification [3]. It performs better on Reuter's data collections. The classification problem consists of a single and multi-labelled problem. The common solution to the multiple labeled problems is to decompose it to independent binary classifier. Binary one is assigned to one of two predefined categories, e.g. positive and negative category. Data mining techniques were applied to text mining and classification by extracting co-occurring terms as descriptive phase from document collection. Effectiveness of text mining using phrases does not show much improvement due to large amount incoming training data for an individual category. Text representation is mainly classified based on text classification. Information Extraction

Information Extraction is the method of retrieving relevant data from the document collections. Set of extraction patterns are the key component of the information extraction. The first step of information extraction is to analyze the unstructured data. Key phrases and the relationship within text are identified by the extraction software. This is done by seeing predefined sequences in text called pattern matching. The software identifies the relationships to provide meaningful information to the user. When large volume of data is used this method is very useful. Topic tracking works based on the document that user views. It also predicts other document interest to the user. Free topic tracking tool is offered by yahoo the keyword can be chose by the user when new related topic are present it will notifies. Topic tracking have limitations. An information filtering system views the incoming document stream and selects the document relevant to one or more of its query. The information filter task helps the user by reducing the irrelevant information and provides relevant information [10]. Traditional information filtering model was based on term based method. Most of them were extracted from positive training document. It was assumed that phrase based approach will provide a better result than the term based method but unfortunately experimental result proved that assumption was incorrect. The reason is that long pattern is less useful and low frequency occurs by longer pattern. Pattern also has inferior properties [11].

3.1 Adaptive Information Filtering

An information filtering system views the incoming document and makes a binary decision to accept or reject that document based on the user profile. Feedback to the system can be provided by the user. The main aim of the adaptive information filtering is to automatically retrieve the data stream to the topic specified by the user [4]. The filtering process is done based on some knowledge that the system has extracted from feedback document.

4. TEXT CATEGORIZATION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save. You are now ready to style your paper. The natural language document is assigned to predefined categories according to their content.

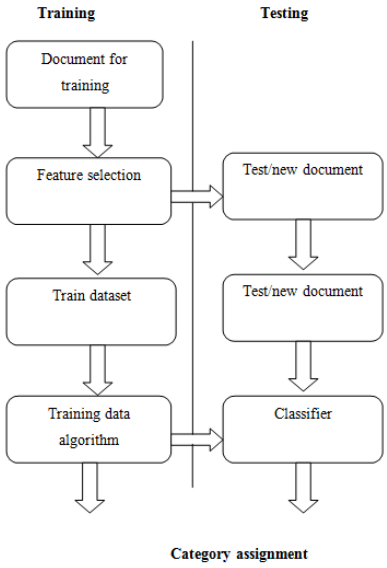

Fig 3: Text Categorization

High dimensionality of feature space is the major characteristics of text classification. algorithm is used to learn the classifiers from labelled documents. The classification is performed automatically on unlabelled documents. Feature space reduction, tokenizing the text, indexing are involved in feature selection. Tokenizing of the text can be done by using term frequency, inverse document frequency and binary representation. There are several categorization methods**.**

4.1 Decision Tree

The manual categorizations of the training document are rebuilt by constructing well distinct true/ false queries in the form of tree structure. Leaves correspond to the category of the document and nodes are questions. The disadvantage of the tree method is "over fitting". A new document can be categorized easily after creating the tree structure. The document is put into the root node and it is allowed to run through the query structure, finally it reaches a certain leaf. The advantage of this structure is that even the people who are unfamiliar about the details of the model can easily interpret the output tree.

4.2 K-nearest Neighbor

The closeness of the document is calculated by Euclidean distance between the vectors. The advantage of this method is simplicity and do not need any resources for training. The disadvantage of this technique is the above average categorization time.

4.3 Bayesian Approach

Naïve and non naïve are the two types of Bayesian approaches. The disadvantage if this approach is that only binary feature vectors can be processed, thus relevant information are possibly abandoned.

4.4 Vector based Method

Centroid algorithm and support vector machine are two types of vector based methods. Centroid algorithm is the simplest algorithm; average feature vector for each category is calculated during the learning stage. By this algorithm new document can be easily categorized. Support vector machine uses negative documents in addition to positive documents. Superior runtime behavior at the time of categorization of new document is the advantage of this method. This is because for each new document only one dot product is calculated [3].

4.5 Keyword-Based Representation

Bag of words method is one of the usual keyword-based representations. It is widely used in the field of the text classification. The advantage of the bag of words is simplicity. Each word that is mined from the document is stored in the vector space. The context of this document is represented by these words which are called as features. The two problems in this approach are synonyms and homonyms. Over fitting and selecting limited number of feature is another disadvantage of this approach.

4.6 Phrase-Based Representation

Keyword-based representation causes the ambiguity problem [15]. To solve this multiple words as features is proposed. Phrases contain more specific content than single word. The advantages are it can automatically discover the hidden semantic sequences of documents under each category, which can benefit the classification accuracy. N-multigram model is related to n-gram model. The disadvantage of the phrase based text representation was pointed out in [11].

4.7 Pattern Based Approach

There are two key factors regarding the efficiency of pattern-based approaches: low frequency and misinterpretation. If the minimum support is decreased a lot of noisy patterns can be found. Misinterpretation means 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.

5. FEATURE SELECTION

Feature selection has vital role in reducing the dimensionality of dataset by removing the irrelevant features. It is done to improve the classification accuracy and reduces the overfitting. Document frequency, information gain, mutual information, term strength and odd ration are some of the measures.

6. PATTERN TAXONOMY MODEL

Pattern taxonomy model is based on pattern methodology. The pattern taxonomy consists of two stages. The first stage

extracts useful phrases from text document. The weight of the term which is occurring in the extracted pattern is then calculated to improve judgments on the new document [13]. The pattern taxonomy model discovered closed sequential patterns in the text document. Patterns are set of terms that frequently appeared in paragraph. In these approaches too many noisy patterns adversely affect the pattern taxonomy model systems. Pattern taxonomy model is more reliable using positive training documents only. All documents are divided into paragraphs. So a given document d has a set of paragraphs.

7. PATTERN DEPLOYING METHOD

The importance of patterns can be estimated by assigning an evaluated value based on one existing weighting function. The same is required for the pattern discovery in the phase document evaluation to find the matched pattern. This is ineffective and takes more time which affects the performance. The drawback is on computational expensiveness by the data mining based methods and unsolved low frequency problem of long patterns. Pattern deploying methods are proposed for the use of discovered knowledge [12]. All discovered patterns are not interesting because some noise patterns are also extracted from the training dataset. Information from the negative example is not exploited during that concept learning. The negative document also contains useful information to identify ambiguous pattern in the concept. It is easier to find the relevant document if the same pattern appears in the positive document [14]. But if the same pattern appears in the negative document it will be difficult. To increase the efficiency it is necessary for a system to exploit ambiguous pattern from the negative examples in order to reduce their influence**.**

8. PATTERN DEPLOYING METHOD

The term weight is different from term based approach in the pattern taxonomy model. In term based approach the evaluation is done based on the term appears in the documents. The term weight is evaluated by the term appearance in the discovered patterns.

9. INNER PATTERN EVOLUTION

This algorithm helps to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. 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 the pattern. A threshold is used to classify documents into relevant or irrelevant categories. In order to reduce the noise, d-patterns are tracked and find out which pattern give rise to such an error [17]. These patterns are offenders. There are two types of offenders *complete conflict offend* and *Partial conflict offender* the idea of updating patterns is explained as follows: Complete conflict offenders are removed from the discovered d-pattern at first. For partial conflict offenders reshuffling of their term support is carried out in order to reduce the effects of noise documents. This algorithm gives the better result and efficient updating of discovered pattern which is extracted from the text document.

10. CONCLUSION

There were lots of data mining techniques from past decades but, the updating of discovered pattern effectively was difficult with those techniques because the long pattern with high specificity lacks in support. Inadequate use of patterns that are extracted also causes performance degradation. The main issue regarding the pattern based approach is low frequency and misinterpretation. The present research pattern taxonomy model which includes pattern evolving and deploying method helps in the updating of useful pattern efficiently and the two issues can be solved. It helps in finding the useful information to the user. The inner pattern evolution outperforms the pattern deploying method.

11. ACKNOWLEDGMENTS

I thanks to my guide who have helped me to do this survey and all others staffs and friends for helping in knowing and giving support for learning the techniques and methods of the text mining.

12. REFERENCES

[1] R. Agrawal and R. Srikant. "Mining sequential patterns." Research Report RJ 9910, IBM Almaden Research Center, San Jose, California, October 1994.

[2] Hye-Chung Kum, Joong Hyuk Chang, and Wei Wang: "Sequential Pattern Mining in Multi-Databases via Multiple Alignments. IEEE Trans on Data Mining Knowledge and Discovery. 12(2-3): 151-180 (2006).

[3] F. Sebastiani. "Machine learning in automated text categorization." ACM Computing 34(1):1–47, 2002.

[4] Klinkenberg, Ralf and Renz, Ingrid. "Adaptive Information Filtering: Learning in the Presence of Concept Drifts." *learning for Text Categorization, Menlo Park, CA, USA*, AAAI Press, pages 33--40 1998

[5] R. Agrawal and R. Srikant. "Fast algorithms mining association rules." *In Proc.of the VLDB Conference, Santiago*, Chile, September 1994.Expanded version available as IBM Research Report RJ9839, June 1994.

[6] Han, J. and Kamber, M. "Data Mining Concepts and Techniques." *3rd edition, University of Illinois at Urbana-Champaign*, Morgan Kanufmann publishers 2006.

[7] C. Borgelt. Sam: "Simple Algorithms for Frequent Item Set Mining". *IFSA/EUSFLAT* 2009 conference- 2009.

[8] J. Han and J. Pei. "Mining frequent patterns by pattern-growth: methodology and implications." *SIGKDD Explore.* Newsl. 2(2):14{20, 2000.

[9] R. Srikant and R. Agrawal. "Mining sequential patterns: Generalizations and performance improvements." In P. M. G. Apers, M. Bouzeghoub, and G. Gardarin, editors,*Proc. 5th Int. Conf. Extending Database Technology.*

[10] M. J. Zaki. Spade: "An efficient algorithm for mining frequent sequences." *Machine Learning*, 42(1-2):31{60, 2001.

[11] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., and Hsu, M.-C. 2000. "Free span: frequent pattern-projected sequential pattern mining." In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 355{359.

[12] S.-T.Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. *IEEE Sixth*

[13] J.Wang and J. Han. BIDE, "Efficient Mining of Frequent Closed Sequences," *Proceedings of the 2004 IEEE International Conference on Data Engineering (ICDE), pp*. 79–90, 2004.

[14] S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" *In iiWAS'05*, pages 1013–1022, 2005.

[15] Li, Yuefeng , Zhong, Ning "Capturing evolving patterns for ontology-based web mining." In Zhong, N. Tirri, & Yao, Y. (Eds.) *IEEE /WIC/ACM International Joint Conference on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*, 20-24 Beijing, China. Sept. 2004.

[16] Y. Li and N. Zhong. "Mining ontology for automatically acquiring web user information needs." IEEE Trans. On Knowledge and Data Engineering, 18(4):554–568, 2006.

[17] S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" *In iiWAS'05*, pages 1013–1022, 2005.

[18] X. Yan, J. Han, and R. Afshar. ColSpan: Mining closed sequential patterns in large datasets. In SDM'03, 2003.