Effective Pattern Matching Approach for Knowledge Discovery Application

Anisha Radhakrishnan, Mathew Kurian

Abstract—Text mining is the technique of retrieving useful information from huge text documents. When patterns are devised we retrieve high quality information. There are several methods for retrieving the information. In term based method polysemy and synonymy was the disadvantage. Pattern based approach also did not provide much efficient result. But this was improved with the pattern evolving and pattern deploying techniques. Automatic categorization can be applied to the pattern based approach so the documents can be categorized.

Index Terms—text mining, pattern mining, pattern evolving, pattern deploying, automatic categorization.

I. INTRODUCTION

Knowledge discovery and data mining is an area that focuses on the different techniques for retrieving the useful data. KDD methodologies became wide spread due to the rapid growth and usage on internet and database. Text mining is one of the recent trends in the data mining. It is similar to the data mining, except data mining tools handles the structured data whereas text mining will work with both unstructured and semi-structured. Text mining can perform much better in companies which consist of huge amount of diverse information [1]. Text mining also became effective as like computing power and this is so relevant today to any number of various businesses. Text mining helps in finding out the patterns that is interesting from the huge database

Humans have ability to differentiate the patterns and apply accordingly but computer does not have this capability and it cannot effortlessly handle the spelling, contextual meaning, patterns etc. In text mining analysis, a large text document is extracted by using key phrases, concepts. Text mining takes collection of documents as inputs and then preprocessing is done. These documents which are preprocessed goes through text analysis phase, techniques and finally the required information is retrieved. Documents of different formats are collected from various sources. These documents are preprocessed to perform some tasks [2].

II. INFORMATION EXTRACTION

The information extraction helps the computer to analyze the unstructured text. The software helps in identifying the key phrases and relation among the text by finding the predefined sequences in the text document called pattern matching. This technology is helpful when large volume of text is used. The key components of IE are set of patterns that are extracted [19]. Topic tracking is performed based on the documents that are view by the user. It helps in predicting the other interesting documents that are of use for the user. But topic tracking have limitations too. The information filtering system selects the document that is relevant by analyzing the incoming document to one or more query. Traditional filtering was performed based on the term based approach and it was extracted mostly for the positive documents. Later pattern based approach was used but it also have inferior properties.

III. TEXT CATEGORIZATION

The documents can be assigned to the predefined categories based on the content. The major characteristics of the text classification are the high dimensionality of the feature space.

The feature selection involves the indexing; tokenizing the text, feature space reduction. There are mainly two approaches in the text categorization knowledge engineering approach and machine learning approach [3]. In knowledge approach the user defines the rules manually.
defined and the text is classified in given categories. The disadvantage of knowledge engineering approach is the bottleneck. The rule has to be manually defined. The machine learning approach it automatically builds an automatic text classifier by learning a set of pre classified documents.

A. Keyword-Based Representation

Bag of words is one of keyword based method that is widely used. Simplicity is the benefit of this approach. The extracted words from the document are stored in the feature space. Synonyms and homonyms are the disadvantage of this approach. Selecting the limited number of features and overfitting are another issue.

B. Phrase-Based Representation

Keyword representation causes the uncertainty issue [15]. Phrases consist of more precise content than single word. It can automatically discover the hidden semantic sequences of each category in documents; this can profit the classification accuracy. N-multigram model is related to n-gram model.

C. Vector Based Method

There are two types of vector based methods centroid algorithm and support vector machine. The simplest one is centroid algorithm. New document can be easily categorized by this algorithm for each category average feature vector is calculated during learning stage. In support vector machine we use negative documents along with the positive documents. Superior runtime behavior is the advantage of this method at the time of categorizing the new documents because dot product is calculated for every new document.

D. Pattern Based Method

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

IV. PATTERN TAXONOMY MODEL

Pattern taxonomy model is on the base of pattern methodology. This has two stages. In the first stage useful phrases are extracted from the text document. The weight of the term which is occurring in the extracted pattern is found out. It is 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. The documents are divided as paragraphs. So a given document d has a set of paragraphs.

A. Closed Sequential Pattern

Sequential pattern included in no other sequential pattern that has the same support exactly is closed sequential pattern [20]. 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 the space deeply [19]. This method avoids the extending of sequence by detecting the extension that has integrated in sequence.

B. 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.

V. 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 knowledge discovered [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.

VI. PATTERN EVOLVING METHOD

In pattern evolving the patterns are updated. The retrieved patterns are shuffled by using shuffling algorithm and the noised patterns are shuffled. There are two types of offenders complete conflict offenders and partial conflict offenders [21]. The complete offenders are removed and partial offenders are shuffled. The patterns that are above the threshold is noised patterns and its support is shrink else the support is grown.

VII. AUTOMATIC CATEGORIZATION

Automatic text categorization is always a significant application and research topic. Text classification is very necessary due to the huge amount of the text documents. Text classification consists of topic based classification and genre based classification. The data’s for the genre classifications are gathered from the news group, bulletin board and broadcast. Text classification is the process of classifying a document in predefined category. An initial dataset is needed in every supervised machine learning task.

REFERENCES


Knowledge And Data Engineering, Vol. 24, No. 1, January 2012

Anisha Radhakrishnan doing her M.Tech in Computer Science and Engineering from the department of Computer Science in Karunya University and received her B.E. from Annamalai University. Has published one international journal.

Mathew Kurian, doing his Ph.D degree in data mining. Previously worked as Software Engineer with Aricent Technologies (2007-09) and currently he is working as Lecturer in Computer Science Department since June 09. He received his B.tech degree from and received M