A COMPARATIVE STUDY ON DIFFERENT TYPES OF EFFECTIVE METHODS IN TEXT MINING: A SURVEY

#1 Inje Bhushan V. and #2 Prof. Mrs. Ujwalapatil

#1 Post Graduate Student M.E (Computer)
Department of Computer Science and Engineering
R.C.Patel Institute of Technology,
Shirpur, DistDhole, Maharashtra, India.

#2 Associate Professor
(Department of Computer Science & engineering)
Department of Computer Science and Engineering
R.C.Patel Institute of Technology,
Shirpur, DistDhole, Maharashtra, India.

ABSTRACT

Textmining is the one of the most resent area for research because of in databases storing information in text form, to extracting information that is the challenging issue to motivate textmining. This survey paper tries to cover the all textmining method that solves these challenges. We presented an exhaustive survey of different pattern mining methods proposed in the literature. Pattern mining methods have been used to analyze this data and identify patterns. Textmining is the discovery by computer for extracting new, previously unknown information and also by automatically extracting information from different written resources.In this survey paper we discuss such successful techniques they gives effectiveness over information retrieval in textmining.

Keywords: Textmining, Information Retrieval, Sequential pattern model, Pattern taxonomy model.

1 INTRODUCTION

Nowadays most of the information in business, industry, government and other institutions are stored in the form of text into databases. This text database contains semi structured data in that they are not only completely unstructured and structured. For example,
a document may contain a few structured fields, such as title, name of authors, date of publication, category, and so on, but also contain some largely unstructured text components, such as abstract and detail content. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research. So that information retrieval techniques [18], such as text indexing methods, have been developed to handle unstructured documents. On other hand in traditional search, the user is typically looking for already known terms and has been written by someone else. The problem is in result appearing all the material that currently is not relevant to your needs in order to find the relevant information. This is the goal of text mining discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining [2] that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

Figure 1. Shows a generic process model for a text mining application [1]. Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

Information Extraction

In computers firstly it analyze unstructured text is to use information extraction [2]. An information extraction technique identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. The technique infers the relationships between all the identified people, places, and time to provide the user with meaningful information. This technology can be very useful when dealing with large volumes of text. Traditional data mining techniques assumes that the information to be “mined” is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases.
Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge as illustrated in Figure 2 [2].

Knowledge discovery

Knowledge discovery [19] and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of Knowledge discovery in databases. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining.

Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. 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. In this paper, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

Text mining is the discovery of interesting knowledge 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 the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models [4], Rough set models [4], Okapi BM25 and SVM [20] based filtering models.

The advantages of term-based methods in term of performance improvement for IR and machine learning. However, term-based 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.
2. METHODS AND MODELS USED IN TEXTMINING

Traditionally there are so many technique was developed to solve the problem in textmining that is nothing but the relevant information retrieval according to user’s requirement. So that research in textmining broadly divides in several terms to find the solution. The list of the datamining technique are often try to overcome the problem , and the techniques likes Association rule mining[8],Sequential pattern mining[16] , close pattern mining[4] , frequent itemset mining[16] ,maximum pattern mining [4], minimum pattern mining[4] .

According to the information retrieval basically there are four methods are used

1) Term Based Method (TBM).
2) Phrase Based Method (PBM).
3) Concept Based Method (CBM).
4) Pattern taxonomy Method(PTM).

There are some more models are used to evaluate and improving the efficiency in textmining like

A. Sequential pattern mining (SPM).
B. Sequential closed pattern mining (SCPM).
C. Frequent itemset mining (NSPM).
D. Frequent closed itemset mining (NSCPM).

The algorithms from the Data Mining community inherited some characteristics from the association rule mining algorithms, and are best suited to work with many (from hundreds of thousands to millions) sequences with relative small length (from 4 to 20). The first algorithms proposed for this task were AprioriAll [2] and GSP[16], from Agrawal and Srikant. Other algorithms like FreeSpan [8], PrefixSpan [4], SPADE [19], CloSpan [18], SPAM [3], were developed afterwards and successively improved the task of find frequent sequence patterns. Algorithms with particular features like, MEMISP [11] which is a memory indexing approach, or SPIRIT [5], which integrates constraints to the mining process through regular expressions, can also be found in literature.

Term Based Method (TBM).

In TBM [3] include efficient computation performance is the advantages are but in other side there are also the limitation in TBM like it occurring polysemy and synonymy problem polysemy mince word having multiple meaning and synonyms mince multiple word having same meaning.

There are some methods based on TBM like

1. Rocchio and probabilistic models [4].
2. Rough set models [4].
3. BM25 and SVM based filtering models [4].

Phrase Based Method [PBM]

In PBM [4], phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include:

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 [4].
Concept Based Method

Text Mining techniques are mostly based on statistical analysis of a word or phrase. The statistical analysis of a term frequency captures the importance of the term without a document only. But two terms can have the same frequency in the same document. But the meaning that one term contributes might be more appropriate than the meaning contributed by the other term. Hence, the terms that capture the semantics of the text should be given more importance. Here, a new concept-based mining is introduced [6].

In Concept-Based Information Retrieval Using Explicit Semantic Analysis[5] in this paper author Concept-based IR using Explicit Semantic Analysis (ESA) makes use of concepts that encompass human world knowledge, encoded into resources such as Wikipedia (from which an ESA model is generated), and that allow intuitive reasoning and analysis. Feature selection is applied to the query concepts to optimize the representation and remove noise and ambiguity.

Pattern Taxonomy Method

Pattern mining has been extensively studied in data mining communities for many years. 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). Here author NingZhonget.al argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance.

In this research work, an effective pattern discovery technique [3] 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. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept-based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

Sequential pattern mining (SPM)

Before going to elaborate term SPM first we see what is Sequence Data? Sequence data is omnipresent. Customer shopping sequences, medical treatment data, and data related to natural disasters, science and engineering processes data, stocks and markets data, telephone calling patterns, weblog click streams, program execution sequences, DNA sequences and gene expression and structures data are some examples of sequence data. A sequential pattern mining algorithm should

A. Find the complete set of patterns, when possible, satisfying the minimum Support (Frequency) threshold,
B. Be highly efficient, scalable, involving only a small number of database scans
C. Be able to incorporate various kinds of user-specific constraints.

There are two major difficulties in sequential pattern mining:
(1) Effectiveness: the mining may return a huge number of patterns, many of which could be uninteresting to users, and
(2) Efficiency: it often takes substantial computational time and space for mining the complete set of sequential patterns in a large sequence database.
### Table 1. Comparative summery of textmining methods

<table>
<thead>
<tr>
<th>Model/ Method</th>
<th>Approach / Algorithm</th>
<th>Author</th>
<th>Parameters</th>
<th>Inference</th>
<th>Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Retrieval</td>
<td>Rocchio[24]</td>
<td>Thorsten Joachims</td>
<td>Models, documents and queries as TF-IDF vectors.</td>
<td>Learning is very fast in this method</td>
<td>Set of documents that are not relevant</td>
</tr>
<tr>
<td>Close Pattern Mining [7]</td>
<td>CHARM[23]</td>
<td>Mohammed J. Zaki et al</td>
<td>set of all frequent closed item-sets</td>
<td>CHARM performed to discover the longest pattern</td>
<td>IBM Almaden, pumsb and pumsb contain census data</td>
</tr>
<tr>
<td></td>
<td>CloSpan[7]</td>
<td>X. Yan, J. Han, and R. Afshar</td>
<td>frequent patterns in the dataset</td>
<td>It mine long sequence for KDD it produces significantly less number of discovered sequences.</td>
<td>Sequence of $s$, Projected DB D8 and min_sup</td>
</tr>
<tr>
<td>Frequent Itemset Mining[13][27]</td>
<td>FPgrowth[27]</td>
<td>C. Borgelt, 2005</td>
<td>prefix tree representation of the given database of transactions</td>
<td>This algorithm can save considerable amounts of memory for storing the transactions.</td>
<td>set of transactions</td>
</tr>
<tr>
<td>Maximal Pattern Mining [4][28]</td>
<td>MaxMiner[28]</td>
<td>Mohammed J. Zaki</td>
<td>real and synthetic datasets</td>
<td>MaxMiner shows good performance on some datasets, which were not used in previous studies</td>
<td>Frequent itemset</td>
</tr>
<tr>
<td></td>
<td>GenMax[22]</td>
<td>Karam Gouda, Mohammed J. Zaki, 2003</td>
<td>frequent items and the frequent 2-itemsets</td>
<td>This algorithm works 2 times faster than other like Mafia PP using dataset “Chess and pumsb”</td>
<td>Dataset is in the vertical &quot;tidset&quot; format</td>
</tr>
<tr>
<td>Pattern Taxonomy [3]</td>
<td>D-Pattern Mining [3]</td>
<td>NingZhong et al 2012</td>
<td>deploying process, which consists of the d-pattern discovery and term support evaluation</td>
<td>The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.</td>
<td>Positive document $D+$, minimum support $min_sup$</td>
</tr>
</tbody>
</table>
CONCLUSIONS

We discussed basics of text mining method. We presented an exhaustive survey of different pattern mining methods proposed in the literature. Pattern mining methods have been used to analyze this data and identify patterns. Such patterns have been used to implement efficient systems that can recommend based on previously observed patterns, help in making predictions, improve usability of systems, detect events and in general help in making strategic product decisions. We envision that the power of Text mining methods like Sequential pattern mining Pattern taxonomy model has not yet been fully exploited. We hope to see many more strong applications of these methods in a variety of domains in the years to come.

REFERENCES

3. Ning Zhong, Yuefeng Li “Effective pattern discovery in text mining” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. , NO.
13. J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), pp. 1-12, 2000.
17. M. Gupta and J. Han “Approaches for Pattern Discovery Using Sequential Data Mining”, 2011 - Information Science Reference.
23. Mohammed J. Zaki and Ching-Jui Hsiao “CHARM: An Efficient Algorithm for Closed Itemset Mining”.
25. Bart Goethals “Frequent Set Mining” Data Mining and Knowledge Discovery Handbook chapter no. 17.
26. Gosta Grahne and Jianfei Zhu “Efficiently Using Prefix-trees in Mining Frequent Itemsets”
28. Mohammed J. Zaki “Mining Closed & Maximal Frequent Itemsets” NSF CAREER Award IIS-0092978, DOE Early Career Award DE-FG02-02ER25538, NSF grant EIA-0103708.