1. INTRODUCTION

Knowledge discovery has become an indispensable phenomenon in recent years due to the rapid increase in digital data. They have attracted lots of attention in academic and scientific circles. Many applications in the real world need such mining of data in order to discover trends or patterns. These trends or patterns lead to business intelligence (BI). Such BI helps in taking well-informed decisions. Many data mining techniques came into existence in the past ten years. They include closed pattern mining, maximum pattern mining, sequential pattern mining, item set mining, and association rule mining. These techniques are developed for data mining algorithms. They are capable of producing huge numbers of patterns. However, how to use those patterns and how to update them in future is the area that needs some more research. Especially in the field of text mining, patterns are discords from text documents. It is a challenging job to use those patterns and also update them. Earlier term-based methods are provided by Information Retrieval (IR) techniques. The term-based methods are classified into rough set models [1], SVM-based models [2] and probability models [3]. All the term-based methods suffer from problems such as synonymy and polysemy. When award has many meanings it is known as polysemy. When multiple words have similar meaning, it is called synonymy. Thus the discovered patterns with term-based techniques have semantic meaning and answering the exact user query is difficult.

For this reason for many years people started believing that phrase-based techniques are better than that of term-based. However, the experiments in the field of data mining [4], [5], [6] have not been proved. The possible reasons include the phrases have less properties pertaining to statistics when compared with terms; frequency of occurrence is low; noisy and redundant phrases are more [6].

Though there are some drawbacks, the sequential patterns became promising alternatives to phrases [7], [8]. The reason for this is that sequential patterns avail required statistics like terms. Pattern Taxonomy Models (PTMs) [8], [9] came into existence to overcome the drawbacks of phrase-based mining approaches. Pattern-based approaches become alternatives but much improvements are not made to make them more effective for text mining. With regard to effectiveness there are two issues. They are misinterpretation and low frequency. When patterns are less frequent, they can’t be used for decision making. When the terms or patterns are misinterpreted, the result will not be reliable. Low frequency can’t have required support. If the support is decreased, the results may not be useful for business decisions.

Over the last many years Information Retrieval (IR) is also used to have many techniques that used features of text documents. They are used to retrieve content from huge amount of documents based on the terms and their weights. The terms may have different weights based on the context as well. There might be semantic meanings that are to be considered in IR. Therefore it is not sufficient to only consider weights of terms for document analysis or evaluation. In this paper we implement a novel pattern discovery technique proposed by Zhong et al. [10]. It first computes specificities of the discovered patterns and then evaluates the weights of terms based on the distribution. Thus it is capable of avoiding misinterpretation problem. Negative training examples influence is also considered by this in order to avoid low-frequency problem. Moreover the ambiguous patterns are updated. This phenomenon is known as pattern evaluation.
Thus the proposed approach improves accuracy of the discovered patterns.

The remainder of this paper is organized as follows. Section II provides review of literature. Section III provides details of the proposed technique. Section IV presents implementation details. Section V provides experimental results while section VI concludes the paper.

2. PRIOR WORK

Textual documents are increasingly added to the World Wide Web and also the electronic databases of organizations. One of the representations which are well known is known as bag of words approach that makes use of keywords. Tf*idf weighting scheme is presented in [11] for representing text. In [12] entropy weighting and global IDF are used for text representation in addition to DFIDF. For the approach bag of words various schemes were developed for weighting [13], [14], [15]. The drawback in the bag of words is that choosing limited number of words is a problem thus it causes over fitting [6]. To reduce number of features other approaches came into existence. They include Odds ratio, Chi-Square, Mutual Information, and Information Gain [4], [6]. Though there are many representations, the choice of representation is based on the requirement, the rules of natural language [6].

Some researchers used phrases instead of words. Unigram and bigram combination is also used in the text categorization process. Phrase based approach is explored in [16]. Data mining techniques are also used as explored in [17]. There was no significant improvement in text mining when phrases are used. It suffered from lower frequency and misinterpretation problems [18]. Some insights were provided on ontology mining which is again term based [19], [20]. In [21] a technique known as pattern evolution was introduced. In data mining communities, pattern mining is extensively used for number of years. Algorithms such as GST [22], SLPMiner [23], SPADE [56] etc. are used for the purpose of data mining. However, finding interesting patterns is still open to anyone to research [25], [26]. Pattern mining is also used in text mining domain. Frequently found items is used text mining for various decisions making applications. Closed sequential patterns are also explored in text mining [9]. In [10] a model known as Pattern Taxonomy model is proposed in order to improve the discovered patterns in text mining. In [27] a two-stage model was developed. The two stages include pattern based methods and term based methods. For text mining “Natural Language Processing” concepts are used. Recently a new model known as concept-based model came into existence [28], [29]. Conceptual Ontology Graph is also explored in order to use semantic knowledge in the discovery of patterns. This model provides effective discrimination between meaningful terms and important terms. In this paper Pattern Taxonomy Model is used for text mining.

3. PATTERN TAXONOMY MODEL

The pattern taxonomy model described briefly here refers to Zhong et al. [10]. The PTM approach assumes that all text documents are converted to paragraphs. Therefore any given document is a set of paragraphs. By using “is a” relation it is possible to structure documents into taxonomy. Consider the following table.

<table>
<thead>
<tr>
<th>Frequent Pattern</th>
<th>Covering Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>{t_3, t_4, t_6}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_3, t_4}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_3, t_6}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_1, t_6}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_3}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_4}</td>
<td>{dp_2, dp_3, dp_4}</td>
</tr>
<tr>
<td>{t_1, t_2}</td>
<td>{dp_1, dp_5, dp_6}</td>
</tr>
<tr>
<td>{t_1}</td>
<td>{dp_1, dp_5, dp_6}</td>
</tr>
<tr>
<td>{t_2}</td>
<td>{dp_1, dp_5, dp_6}</td>
</tr>
<tr>
<td>{t_6}</td>
<td>{dp_2, dp_3, dp_4, dp_5, dp_6}</td>
</tr>
</tbody>
</table>

As can be seen in table 1, frequent patterns are shows in the left column while the right column shows the documents in which these patterns exist. This is the based to structure pattern taxonomy. The constructed pattern taxonomy for the given values in table 1 is as shown in fig. 1.

![Fig. 1 – Pattern taxonomy (excerpt from [10])](http://www.ijret.org)
4. PROTOTYPE IMPLEMENTATION

The pattern discovery technique proposed by Zhong et al. [10] has been implemented by us using Java programming language. The environment used for the implementation include a PC with 4GB RAM, Core 2 Dual processor. Operating system used is Windows and the IDE is Net Beans. Java SWING API is used to build GUI (Graphical User Interface). The main UI of the application is as shown in fig. 2.

As seen in fig. 2 the application facilitates preprocessing before actual discovery of patterns. The selected dataset is shown in text area. Before proceeding further, the text needs to be preprocessed for operations like removal of stop words and stemming. On choosing preprocessing, the UI as shown in fig. 3 is rendered.

As can be seen in fig. 3, there is provision for stop words removal and stemming. These two are the fundamental preprocessing operations required before actually processing the text documents. The PIM button helps to build a pattern taxonomy model. The discovered patterns are shown in fig. 4.

As can be seen in fig. 5, the terms for each discovered pattern are presented. The pattern1 has terms such as t6, t7, t8 and t9. This way the terms are shown for all discovered patterns. Groups of terms involved in different patterns are extracted and presented in fig. 6.
As can be seen in fig. 6, the discovered patterns distribution is shown. The terms are grouped according to the patterns to which they belong to. As can be observed in fig. 6, the results show the terms which do not belong to any patterns. Such terms are pruned and the results are presented in fig. 7.

![Fig. 7 – Final results showing distribution of terms and patterns](image)

As can be seen in fig. 7, the results reveal the distribution of terms and corresponding discovered patterns.

**CONCLUSIONS**

Data mining techniques have been around for long time. The techniques used to discover knowledge include sequential pattern mining, frequent item set mining, closed pattern mining and maximum pattern mining. These data mining techniques are not useful for text mining. This is due to lack of high specificity of discovered patterns. Not all frequent patterns discovered by mining algorithm are useful. Moreover then can be misinterpreted to make the problem worse. To overcome the problems of misinterpretation and low frequency, we proposed an effective pattern discovery. Pattern deploying and evolving are the two parts in the proposed technique. The empirical results revealed that the proposed technique is effective.

**REFERENCES**


