International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Pattern Discovery Techniques for the Text Mining and its Applications

Minakshi R. Shinde[1], Parmeet C. Gill[2]

[1, 2] BAMU University, Marathwada Institute of Technology, Aurangabad, 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. This paper include different 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, text classification, pattern mining, pattern evolving, data mining.

1. Introduction

Text mining is the discovery of interesting knowledge in text documents. It is challenging issue to find accurate knowledge in text documents to help users to find what they want. 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 effectively use and update discovered patterns and apply it to field of text mining [2][9]. Data mining is therefore an essential step in the process of knowledge discovery in databases, which means data mining is having all methods of knowledge discovery process and presenting modeling phase that is application of methods and algorithm for calculation of search pattern or models. 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 [4]. With a large number of patterns generated by using the data mining approaches, how to effectively exploit these patterns is still an open research issue.

Text mining is the technique that helps users find useful information from a large amount of digital text data [3]. It is therefore crucial that a good text mining model should retrieve the information that users require with relevant efficiency. Traditional Information Retrieval (IR) has the same objective of automatically retrieving as many relevant documents as possible whilst filtering out irrelevant documents at the same time. However, IR-based systems do not adequately provide users with what they really need. Many text mining methods have been developed in order to achieve the goal of retrieving for information for users. 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. The process of knowledge discovery may consist as following:

- Data Selection
- Data Processing
- Data Transaction
- Pattern Discovery
- Pattern Evaluation.

Text mining is also called as knowledge discovery in databases because, we frequently find in literature text mining as a process with series of partial steps among other things also information extraction as well as the use of data mining. When we analyze data in knowledge discovery in databases is aims of finding hidden patterns as well as connections in those data. While the ability to search for keywords or phrases in a collection are now widespread such search only marginally supports discovery because the user has to decide on the words to look for. On the other hand, text mining results can suggest “interesting” patterns to look at, and the user can then accept or reject these patterns as interesting. In this paper we discussed pattern taxonomy model which extracting descriptive frequent patterns by pruning the meaningless ones. Patterns are sorted based on their reparations.

2. Background

2.1 Text Mining

Text mining is nothing but data mining, as the application of algorithm as well as methods from the machine learning and statistics to text with goal of finding useful pattern, Whereas data mining belongs in the corporate world because that's where most databases are, text mining promises to move machine learning technology out of the companies and into the home" as an increasingly necessary Internet adjunct (Witten & Frank, 2000) – i.e., as "web data mining" (Hearst, 1997). Laender, Ribeiro-Neto, da Silva, and Teixeira (2001) provide a current review of web data extraction tools.

Text mining is also referred to as text data mining, roughly equivalent to text analytics; it refers to process of deriving high quality information form text. and high quality of information is derived through devising of patterns. Text analysis involves information retrieval, lexical analysis, word frequency distributions, pattern recognition, information extraction, and data mining techniques including link and association analysis, visualization to turn text into data for analysis via..Natural language processing and analytical methods. On other hand we called -Text mining is a variation on field called data mining that tries to find interesting patterns from large datasets. This is a concept of text mining describe in this section.

2.2 Pattern Discovery

The pattern used as a word or phase that is extracted from the text document. There are numbers of patterns which may be discovered from a text document, but not all of them are interesting. Only those evaluated to be interesting in some manner are viewed as useful knowledge. It is midfield task between association rule mining and inductive learning. It aims at finding patterns in labelled data that are descriptive. A system may encounter a problem where a discovered pattern is not interesting a user. Such patterns are not qualified as knowledge. Therefore, a knowledge discovery system should have the capability of deciding whether a pattern is interesting enough to form knowledge in the current context.

2.3 Pattern Taxonomy

Pattern can be structured into taxonomy-used knowledge discovery model is developed towards applying data mining techniques to practical text mining applications. Knowledge Discovery in Databases (KDD) can be referred to as the term of data mining which aims for discovering interesting patterns or trends from a database. In particular, a process of turning low-level data into high-level knowledge is denoted as KDD. The concept of KDD process is the data mining for extracting patterns from data. We focus on development of knowledge discovery model to effectively use & update discovered patterns and apply it to the field of text mining.

3. Techniques

Researchers in the text mining community have been trying to apply many techniques or methods such as rule-based, knowledge based, statistical and machine-learning-based approaches. However, the fundamental methods for text mining are natural language processing (NLP) and information extraction (IE) techniques. The former technique focuses on text processing while the latter focuses on extracting information from actual texts. Once extracted, the information can then be stored in databases to be queried, data mined, summarized in a natural language and so on. The use of natural language processing techniques enables text mining tools to get closer to the semantics of a text source [12]. This is important, especially when the text mining tool is expected to discover knowledge from texts.

3.1 Natural Language Processing (NLP)

NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic reliazer to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. The most well known NLG application is machine translation system. The system analyzes texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least of one the following components; tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis. In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark. Morphological or lexical analysis is a process where each word is tagged with its part of speech. The complexity arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used [11].

Semantic analysis is a process of translating a syntactic structure of a sentence into a semantic representation that is precise and unambiguous representation of the meaning expressed by the sentence. A semantic representation allows a system to perform an appropriate task in its application domain. The semantic representation is in a formally specified language. The language has expressions for real world objects, events, concepts, their properties and relationships, and so on. Semantic interpretation can be conducted in two steps: context independent interpretation and context interpretation. Context independent interpretation concerns what words mean and how these meanings combine in sentences to form sentence meanings. Context interpretation concerns how the context affects the interpretation of the sentence. The context of the sentence includes the situation in which the sentence is used, the immediately preceding sentences, and so on.

3.2 Information Extraction (IE)

IE involves directly with text mining process by extracting useful information from the texts. IE deals with the extraction of specified entities, events and relationships from unrestricted text sources. IE can be described as the creation of a structured representation of selected information drawn from texts. In IE natural language texts

are mapped to be predefine, structured representation, or templates, which, when it is filled, represent an extract of key information from the original text [13], [14]. The goal is to find specific data or information in natural language texts. Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Using IE approach, events, facts and entities are extracted and stored into a structured database. Then data mining techniques can be applied to the data for discovering new knowledge. Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications. According to [15] and [16] typical IE are developed using the following three steps:

- **text pre-processing;** whose level ranges from text segmentation into sentences and sentences into tokens, and from tokens into full syntactic analysis
- **rule selection;** the extraction rules are associated with triggers (e.g. keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;
- **rule application**, which checks the conditions of the selected rules and fill in the form according to the conclusions of the matching rules.

3.3 Text Summarization

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization.

For example, summarization tools may extract the sentences which follow the key phrase "in conclusion", after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic. If organizations, medical personnel, or other researchers were given hundreds of documents that addressed their topic of interest, then summarization tools could be used to reduce the time spent sorting through the material. Individuals would be able to more quickly assess the relevance of the information to the topic they are interested in.

An automatic summarization [17] process can be divided into three steps: (1)In the preprocessing step a structured representation of the original text is obtained; (2) In the processing step an algorithm must transform the text structure into a summary structure; and (3) In the generation step the final summary is obtained from the summary structure. The methods of summarization can be classified, in terms of the level in the linguistic space, in two broad groups: (a) shallow approaches, which are restricted to the syntactic level of representation and try to extract salient parts of the text in a convenient way; and (b) deeper approaches, which assume a semantics level of representation of the original text and involve linguistic processing at some level. In the first approach the aim of the preprocessing step is to reduce the dimensionality of the representation space, and it normally includes: (i) stop-word elimination –common words with no semantics and which do not aggregate relevant information to the task (e.g., "the", "a") are eliminated; (ii) case folding: consists of converting all the characters to the same kind of letter case - either upper case or lower case; (iii) stemming: syntactically-similar words, such as plurals, verbal variations, etc. are considered similar; the purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics. A frequently employed text model is the vector model. After the preprocessing step each text element –a sentence in the case of text summarization – is considered as a N-dimensional vector. So it is possible to use some metric in this space to measure similarity between text elements. The most employed metric is the cosine measure, defined as $\cos q = (\langle x.y \rangle) / (|x| \cdot |y|)$ for vectors x and y, where $(<,>)$ indicates the scalar product, and $|x|$ indicates the module of x. Therefore maximum similarity corresponds to $\cos q = 1$, whereas $\cos q = 0$ indicates total discrepancy between the text elements. To implement text summarization based on fuzzy logic, MATLAB is usually used since it is possible to simulate fuzzy logic in this software. Select characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system. Then, all the rules needed for summarization are entered in the knowledge base of this system. Afterward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.

3.4 Categorization

Categorization involves identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not attempt to process the actual information as information extraction does. Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic. As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first. Categorization can be used in a number of application domains. Many businesses and industries provide customer support or have to answer questions on a variety of topics from their customers. If they can use categorization schemes to classify the documents by topic, then customers or end users will be able to access the information they seek much more readily. The goal of text categorization is to classify a set of documents into a fixed number of predefined categories. Each document may belong to more than one class

4. Text Mining Application

The main Text Mining applications [18] are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administration and legal documents.
- Pharmaceutical and research companies and healthcare.
- National Security /Intelligence
- Natural Language/Semantic Toolkit or Service
- Publishing
- Automated ad placement

The sectors analyzed are characterized by a fair variety in the applications being experimented. however, it is possible to identify some sectorial specifications in the use of TM, linked to the type of production and the objectives of the knowledge management leading them to use TM. The publishing sector, for example, is marked by prevalence of Extraction Transformation Loading applications for the cataloguing, producing and the optimization of the information retrieval. In the banking and insurance sectors, on the other hand, CRM applications are prevalent and aimed at improving the management of customer communication, by automatic systems of message re-routing and with applications supporting the search engines asking questions in natural language. In the medical and pharmaceutical sectors, applications of Competitive Intelligence and Technology Watch are widespread for the analysis, classification and extraction of information from articles, scientific abstracts and patents. A sector in which several types of applications are widely used is that of the telecommunications and service companies: the most important objectives of these industries are that all applications find an answer, from market analysis to human resources management, from spelling correction to customer opinion survey.

5. Conclusion

This paper has presented an overview techniques, applications in text mining. The focus has been given on fundamental methods for conducting text mining. The methods include natural language processing and information extraction. The purpose of this section is to give an overview to a reader on how text mining systems can be used in real life.

Acknowledgements

The authors express gratitude to Principal, Head of Department (CSE) **Dr. Radhakrishna Naik.** Marathwada Institute of Technology College of Engineering, Aurangabad, and Maharashtra India. They also express their sincere thanks all the faculty members of CSE Department MIT College of Engineering, Aurangabad, and Maharashtra, India for their constant support and enthusiasm.

References

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.
[2] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
[3] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.
[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.
[5] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[6] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[7] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc.

14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.
[8] A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.
[9] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining",vol.24,No.1,Jan.2012.
[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] S.Jusoh and H.M. Alfawareh, "Natural language interface for online sales," in *Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007)*. Malaysia: IEEE, November 2007, pp. 224–228.
[12] S. Jusoh and H. M. Alfawareh, "Agent-based knowledge mining architecture," in *Proceedings of the 2009 International Conference on Computer Engineering and Applications*, IACSIT. Manila, Phillipphines: World Academic Union, June 2009, pp. 602–606.
[13] R. Rao, "From unstructured data to actionable intelligence," in *Proceedings of the IEEE Computer Society*, 2003.
[14] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in *Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference*, 2000.
[15] R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson, "FASTUS: A cascaded finite-state transducer for extraction information from natural language text," in *Finite States Devices for Natural Language Processing*, E. Roche and Y. Schabes, Eds., 1997, pp. 383– 406.
[16] J. Cowie and Y. Wilks, *Information extraction*, New York, 2000.
[17] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy and Asghar Tajoddin (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer society, 347-352.
[18] Sergio Bolasco , Alessio Canzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2002), "Understanding Text Mining:a Pragmatic Approach", Roam, Italy

Author Profile

Minakshi R.Shinde received the B.E degree in Information Technology from HI-TECH Institute of Technology from Aurangabad in 2011 at present appearing the M.E degree in Computer Science and Engineering department at Marathwada Institute of Technology, Aurangabad. Her research interest in Pattern discovery techniques for the text mining and its application