A SURVEY OF PATTERN DISCOVERY METHODS FOR TEXT MINING

Mrs.Neha Sangram Patil
Assistant Professor, AISSMS IOIT, Pune

Abstract— Too many data mining methods are used to mine useful patterns from text documents. Accurate and exact knowledge in the text documents is needed for user to find what they wish. Text mining is discovery of interesting knowledge in text documents. Using and updating discovered patterns is still open research issue. Lot of term based methods are present but disadvantage with these methods is that they suffer from problem of polysemy and synonymy. To overcome these disadvantages pattern mining methods are suggested. Because of low frequency and pattern misinterpretation problem, pattern mining methods are not proved to be better than term based methods. An effective pattern discovery technique is suggested which includes processes of pattern deploying to overcome misinterpretation problem and pattern evolving to overcome low frequency problem. This improves performance of modifying discovered patterns and finding interesting and relevant information in text mining.

Index Terms— Text mining, pattern evolving, pattern mining, text classification.

I. INTRODUCTION

The need of storing information and volumes of data in computers has slowly leaded to data mining. Data mining perform real time analysis on data which is to be used for further processing. Actual work of data mining is to find interesting patterns which are present in data. These patterns are helpful for user and contain hidden information. These methods consist of frequent item mining, association rule mining, closed pattern mining and sequential pattern mining. data mining approaches generate varieties of patterns but how to effectively use and update these discovered patterns is an open research issue. Here motto is to do pattern discovery to build the information discovery model to effectively utilize the discovered patterns and apply it to text mining. Text mining is discovery of information in text documents. Information Retrieval gives us many term based methods like BM25, Rocchio's and probabilistic model and support vector machine (SVM). The term based methods suggests well performance as they support mature theories for term weighing which come from IR communities. But term based methods suffer from the issues of synonymy and polysemy. There is hypothesis over the years that phrase based approaches may perform higher than the term based ones because phrases could carry additional semantics like information. pattern mining methods are evolved to overcome phrase based approaches which uses closed sequential patterns .But pattern based approaches also not proved to be significant improvement over term based approaches because of low frequency and pattern misinterpretation problem. To overcome these problems effective pattern discovery techniques are suggested and uses pattern deploying and pattern evolving process. For a specific topic large patterns are more related to topic but with low support. If minimum term support is decreased lot of noisy patterns would be discovered. To overcome low frequency problem suggested technique uses pattern evolution technique. This technique considers the influence of patterns from the negative training examples that help to identify ambiguous patterns and try to overcome their influence for the low-frequency drawback. The process of updating ambiguous patterns is nothing but Pattern evolution. Misinterpretation problem measures confidence and support turn out to be not suitable in using discovered patterns to give answer to what user want. Terms are very important feature in text document according to IR theory. Terms with larger weight scheme are more general terms because frequently used in both irrelevant and relevant information. But it is not sufficient to evaluate term weights based on their distribution in documents for a specific topic. To overcome misinterpretation problem, effective pattern discovery technique use pattern deploying process to weights the term according to its specificity in discovered pattern instead in document.

II. Related work

In the past many texts representation methods are present. Most of these data mining methods adopt term based approach. Ordering of terms in document is less important for many tasks. The main drawback in Bag of words approach is that relationship among words. The relationship cannot be reflected. Next problem is semantic ambiguity while considering single words which can be defined as synonyms: two words share same meaning and homonym: a word with having two or more meanings. Support vector machine is popular learning method. It is well suited for text categorization. SVM uses properties of text such as high dimensional space, few irreverent vectors and sparse instance vector. Having high dimensional feature space SVM removes need for feature selection and also not require parameter tuning. Without giving threshold for defining decision rule for class membership,TFIDF classifier algorithm returns value as ranking of a given document. Rocchio's algorithm is widely used learning method in information retrieval approach. For categorization of text Probabilistic analysis of Rocchio's method is used. In Rocchio's algorithm heuristics are used for similarity matrix word and weighing scheme. Probabilistic variant of Rocchio's algorithm is best over heuristic algorithm. There are number of drawbacks with Rocchio's classifier that low classification accuracy. Heuristic components of the algorithm give many design choices and there is little direction while applying this algorithm to new domain. Algorithm developed and also optimized for relevance feedback in information retrieval and no guidelines are given which heuristics work best for text categorization. Term based methods has efficient computational performance and mature theories of term weighing. To overcome drawback of semantic ambiguity problem nothing but keyword-based representation, use of multiple words or phrases as features is suggested. In general phrases carry more specific meaning than individual words. To overcome problems with phrase based approaches pattern based methods are used. In Automatic Pattern-Taxonomy Extraction for Web Mining gives model for discovering the phrases i.e. frequent sequential patterns by pruning the meaningless phrases. This paper shows that pattern based methods outperforms keyword based method. Sequential patterns means phases processed by sophisticated data processing to make web system more efficient by finding the relevant documents according to user need. This is done by showing relationship between the discovered phrases. The discovered knowledge is represented by correct phrase taxonomy. This approach present PTM pattern-based model to represent text document. This structure is a tree-like that represent the relationship in extracted pattern from text collection.

III. Techniques of Pattern Discovery

The process or practice of examining large collections of written resources in order to generate new information is nothing but Text mining. The goal of text mining is to discover relevant information in **text** by transforming the text into data that can be used for further analysis. Text mining achieves this through the use of a variety of analysis methodologies. Natural language processing (NLP) is one of them. In text mining the main scope is to discover relevant information that is possibly unknown and hidden in the context of other information.

1. Natural language processing (NLP)

Natural language processing (or NLP) is a component of text mining that performs a special kind of linguistic analysis that essentially helps a machine "read" text. NLP uses a variety of methodologies to decipher the ambiguities in human language, including the following: automatic summarization, part-of-speech tagging, disambiguation, entity extraction and relations extraction, as well as disambiguation and natural language understanding and recognition.

To work, any natural language processing software needs a consistent knowledge base such as a detailed thesaurus, a lexicon of words, a data set for linguistic and grammatical rules, an ontology and up-to-date entities. NLP software is a "shadow" process running in the background of many common applications such as the personal assistant features in smart phones, translation software and in self-service phone banking applications. Text mining and NLP are commonly used together for different purposes, and one of most common applications is social media monitoring

2. Information Extraction

The problem of text mining, i.e. discovering useful knowledge from unstructured or semi-structured text, is attracting increasing attention. A new framework for text mining based on the integration of Information Extraction (IE) and Knowledge Discovery from Databases (KDD), a.k.a. data mining. KDD and IE are both topics of significant recent interest. KDD considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. IE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from free text. However, there has been little if any research exploring the interaction between these two important areas.

Traditional data mining 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 . Information extraction can play an obvious role in text mining as illustrated.

3. Text Summarization

Text summarization a process of reducing the size of the original document while preserving its information content and its summary is less than half of the main text. Summarization has been viewed as a two step process. The first step is the extraction of important concepts from the source text by building an intermediate representation of some sort. The second step uses this intermediate representation to generate a summary. News blaster is a good example of a text summarizer, that helps users find the news that is of most interest to them. The system automatically collects, cluster, categorizes, and summarizes news from several sites on the web on a daily basis. A summarization machine can be viewed as a system which accepts either a single document or multiple documents or a query as an input and produces a abstract or extract summary.

4. Categorization

Categorization involves identifying the main parts 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 will not attempt to process the actual information as information extraction does. Instead 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

IV. Conclusion

Lot of data mining techniques have been sugggested to discover patterns. But using these patterns and or discovered knowledge in the field of text mining is ineffective. By using pattern discovery technique effective patterns are discovered. This survey give idea about technique to overcome problems of low-frequency and pattern misinterpretations. The purpose of this article is to give an overview to a reader on how text mining systems can be used in real life.

V. References

[1] Wu, Sheng-Tang, Yuefeng Li and Yue Xu "Deploying approaches for pattern refinement in text mining", Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006 October 2011.

[2] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell'Informazione, 2000.

[3] Wu, Sheng-Tang, et al \Automatic pattern-taxonomy extraction for web mining", Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on. IEEE, 2004.

[4] Ahonen, Helena, et al. "Applying data mining techniques for descriptive phrase extraction in digital document collections", Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 1998.

[5] Li, Yuefeng, Abdulmohsen Algarni "Mining positive and negative patterns for relevance feature discovery", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[6] Zhong, Ning, Yuefeng Li and Sheng-Tang Wu \Effective pattern discovery for text mining", Knowledge and Data Engineering, IEEE Trans-actions on 24.1 (2012): 30-44.

[7] Kavitha Murugeshan, Neeraj RK "Discovering Patterns to Produce Effective Output through Text Mining Using Naive Bayesian Algorithm", International Journal of Science and Applied Information Technology,Volume 1, No.3, July { August 2012}.

[8] Ahonen, Helena, et al. "Applying data mining techniques for descriptive phrase extraction in digital document collections", Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 1998.

[9] Yin, Shiqun, et al. \Research and implement of classification algorithm on web text mining", Semantics, Knowledge and Grid, Third International Conference on. IEEE, 2007.

[10] Albathan, Mubarak, Yuefeng Li and Abdulmohsen Algarni "Using Patterns Co-occurrence Matrix for Cleaning Closed Sequential Patterns for Text Mining", Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2012.