IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-3 E-ISSN: 2347-2693

Comparative Study on Information Retrieval Approaches for Text Mining

Vishakha D. Bhope* and Sachin N. Deshmukh
[1,2]Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

www.ijcseonline.org

Received: Mar/01/2015 Revised: Mar/07/2015 Accepted: Mar/22/2015 Published: Mar/31/2015

***Abstract*—** Text mining is the process of extracting information form unstructured to structured text data. The challenging issue in text mining is to extract user required information in efficient manner. To perform this task various data mining methods are used in which the text document analyzed on the basis of term, phrase, concept and pattern. This paper studies the text representation methods and basic term weighing schemes. Ruled-based Phrase Extraction method and Sequential Pattern mining method are discussed to improve the system performance for finding relevant and interesting information.

Keywords— Text Mining, Text Representaion, Rule based Phrase Extraction, Sequential Pattern Mining

I. INTRODUCTION

A great amount of digital data is now available in science, industry, business and many other areas due to advances in computerization and digitization techniques. In order to understand, analyze and eventually make use of huge amount of data the knowledge discovery and data mining techniques are used. Data mining is the essential step in process of knowledge discovery in databases. There are various techniques present for performing different knowledge tasks such as association rule mining, sequential pattern mining, frequent itemset mining and so on[1].

Text mining is the process of extracting meaningful information from text data. Information can be extracted to make analysis of data using various data mining (Statistics and Machine Learning) algorithms. Text Mining comprises of various functions which include Information Retrieval, Information Extraction (IE), Categorization, Summarization, Clustering and Question and Answers. To evaluate these functions efficiently techniques used are, lexical analysis, pattern recognition, word frequency distributions and data mining techniques including association analysis, visualization to turn text into data for analysis via natural language processing methods[2].

Text mining process starts with a document collection from various resources. Text retrieval techniques are used to retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, giving a rich amount of knowledge for the user of that system.

Information Retrieval typically deals with crawling, retrieving documents, indexing documents, parsing. Information Retrieval is more to do with the search engines, concerning mostly about document retrieval (through its text most of the times). For large collection of text data simply scanning is not feasible. Therefore a set of representative keywords must be selected and attached to documents. It is essential that a good text mining model should retrieve the information that meets users' needs within a relatively efficient time frame.

The rest of this paper is structured as follows: Section II presents basic four text representation approaches such as term-based, phrase-based, concept-based and pattern-based approach. Section III provides the basic term weighting schemes to select the terms with valuable information. Section IV presents the two algorithms, phrasal based and pattern based for text retrieval Section V discusses the comparative analysis for the algorithms.

II. TEXT REPRESENTATION APPROACHES

Traditionally there are so many techniques available to solve the problems of text mining for relevant information retrieval as per user's requirement. In text mining functions such as information extraction, categorization, text document analyzed on the basis of term, phrase, concept and pattern. Basically there are four methods for text representation discussed below.

A. Term Based Method

Term is a word in a document having semantic meaning. In term based method on the basis of terms the document is analyzed for efficient computational performance. These

Corresponding Author: *Vishakha Bhope, vishakhabhope@gmail.com*
Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, India

techniques are developed from the information retrieval and machine learning communities.

The Bag of Words approach is typical term-based representation in information retrieval domain. The Bag of Word approach has been widely used because of its simplicity. The Fig. 1 illustrates the Bag of Word technique [3]. As shown in the figure, each word is retrieved and stored in vector space alone with its frequency. The problem with the bag of words scheme is how to select the limited number of features among the enormous set of words or terms in order to increase the systems efficiency and avoid overfitting. To reduce the number of features in Bag of Word scheme, many dimensionality reduction techniques have been proposed using feature selection techniques [4].

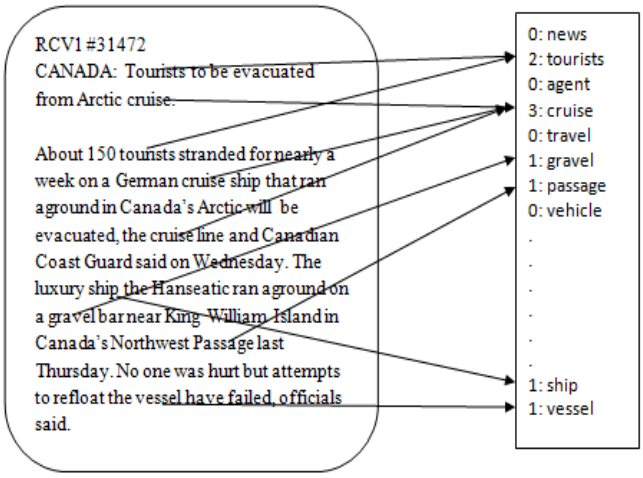

Fig.1 Bag of Words representation using word frequency

Problem with term based approach is the semantic ambiguity which can be divided into synonymy and polysemy, where synonymy is multiple words having the same meaning and polysemy means a word has multiple meanings. Therefore for answering what users want, the semantic meaning of many discovered terms is uncertain [1].

B. Phrase based Method

To avoid the semantic ambiguity problem of term based methods new method is developed using multiple words (i.e. phrase) as a feature. For example, "mining" and "data mining". This approach can find accurate discrimination of the content in documents because in term-based approach a single word is rarely specific. It is considered that the phrase based approaches could perform better than term based ones as more semantic information is carried by phrase than single term. But from several experiments results shown that the phrase based method is not superior to the term based method. Although phrases are less ambiguous and have more brief meaning than individual term, the likely reasons for the dispiriting performance include: 1) phrases have inferior statistical properties to terms, 2) low frequency of occurrence, and 3) large numbers of redundant and noisy phrases are present among them [4].

There are four categories of phrase extraction:

1. Co-occurring terms

If a term appears repeatedly with a particular subset of frequent terms, then the term is likely to have an important meaning [5]. Co-occurrence of terms and frequent terms are counted using term weighting approaches. To find co-occurring phrases, a set of word sequences or text phrases is created for each single document. A set of phrases can be regarded as a content descriptor that should distinguish the document from other documents in the collection [6].

2. Episodes

Episode rules and episodes are a modification of the concept of association rules and frequent sets applied to sequential data [7]. Text episodes can be a sequence of pairs (feature vector, index), where feature vector consist of an ordered set of features such as a word, grammatical feature, punctuation mark or other special character and index is about the position of the word in the sequence. An episode rule is the conditional probability that a certain episode occurs, given that a subepisode has occurred [8].

3. Noun phrases

A text document may contain proper nouns, dates, numbers, punctuations etc. Therefore an appropriate token recognition technique is necessary for efficient text representation. The lexical analysis phase is significant which includes important processes like name recognition, punctuation handling and so on. Syntactic analysis phase involves identification of phrases in a text document. In some techniques Context Free Grammar (CFG) has been used to detect phrases from a text document. A partial parsing is used in which parser aims to find an appropriate phrases from the input text rather than by analyzing the syntactic structure of the text [9].

4. Keyphrases

Many journals ask their authors to provide a list of keywords for their articles. These keywords are called keyphrases, because they are often phrases of two or more words, rather than single words. Keyphrases serves multiple goals like summarization, indexing and query for search engines. The commercial software products are available such as MS office Word uses automatic keyphrase extraction algorithms [10]. There are three keyphrase extraction methods, TFIDF, KEA (Automatic Keyword Extraction), and Keyterm [11].

C. Concept Based Method

Concept-based mining method consists of concept-based analysis of terms and a concept-based similarity measure. The term which has a semantic role in the sentence, is called

a concept. Concepts can be either words or phrases, depends on the semantic structure of the sentence. Conceptual Term Frequency (CTF) is a similarity measure between documents based on a combination of concept-based term analysis similarity within a sentence and concept-based term analysis similarity within a document [12]. The concept based term analyzer algorithm describes how to calculate term frequency and CTF of matched concept in document. Concept-based model can effectively differentiate between non important terms and meaningful terms which describe a sentence meaning [13].

D. *Pattern Taxonomy Method*

Instead of using the phrase-based and term-based methods the efficient way for information retrieval is the pattern-based approach which contains frequent sequential patterns, because sequential patterns have good statistical properties like terms. To overcome the disadvantages of phrase based approaches, pattern taxonomy models have been proposed [14]. In the pattern taxonomy the semantic information will be used to improve the performance by using closed patterns in text mining. Sequential pattern mining concerned with finding relevant patterns in text dataset where values are delivered in sequence. This method uses the concept of closed sequential patterns and pruning non-closed sequential patterns. There are two phases for using the pattern based models in text mining, first is how to discover useful patterns and other is how to utilize those patterns to improve system's performance [15].

III. TERM WEIGHTING SCHEMES

A. *Term Frequency (TF):* The evolution of term weights is based on distribution of term in document. The term frequency TF (t, d) is number of times term t occurs in document d. The TF is calculated to get the term's significance within document.

B. *Document Frequency (DF):* The document frequency DF(t) is number of documents when term t occurs at least once.

C. *Inverse Document Frequency (IDF):* IDF is calculated to get the specificity of terms in a set of documents. Formula for IDF can be expressed as,

$$IDF(t) = log \frac{|D|}{DF(t)} \quad (1)$$

Where |D| is the set of all documents and DF(t) is document frequency. The inverse document frequency of term is low if it occurs in many documents and is highest if term occurs only in one document.

D. *Term Frequency- Inverse Document Frequency (TFIDF):* TFIDF is most widely used measure; it is a combination of term frequency statistics and the specificity statistics (DF) to a term,

$$TFIDF(t) = TF(d,t) \times IDF(t) \quad (2)$$

IV. ALGORITHMS FOR TEXT RETRIEVAL

A. *A Rule-based Phrase Extraction Algorithm*

In a rule-based algorithm keyphrases and keywords are considered as indexing terms. This is a noun phrase extraction algorithm. The extraction of key-phrases from text documents is based on a process of partial parsing. The partial parsing method uses Context Free Grammar (CFG) techniques for extracting phrases from a text document. This algorithm includes two modules [9]:

1. Name Recognition

Proper nouns, currency symbols and dates frequently occur, these names are identified by a set of patterns that are stated in terms of part-of-speech and syntactic features. Name recognition has been researched widely in recent years

2. Syntactic analysis

Depending upon the type of document dataset, and the indexing method used, the size of stopword list keeps varying. A word that appears to be a stopword in isolation forms a meaningful phrase in association with other terms. For example, in "man of the match" phrase all the four words are very common words therefore these words have lower importance in a term-based indexing method. However, when these words appear together in the form of a phrase as "man of the match", it is quite significant. Therefore, this rule-based approach has been developed to identify stopwords. Instead of maintaining a list of stop words, words are grouped into three categories:

a. Content words: words those are meaningful even if they appear alone. For instance, win, appointed, injury, etc.

b. Function words: words those are useful when they appear within a phrase, but not in isolation. For instance, of, the, etc.

c. Stop words: words those are less significant even if they appear in a phrase. For instance, then, they, etc.

This tagging information is stored in the lexical database with their lexical features. This algorithm reads the text document one word at a time and check the word in lexical database to decide the part-of-speech (POS) tagging. If the word is found in lexical database, then it takes the complete phrase by the syntactic analysis process, and if the word is part of a name, then it takes the complete name using the name recognition module. If the word is not available in current dictionary i.e. unknown, then using the POS tag details new word can be added to the dictionary.

The phrase is identified by POS information and CFG rules. Each word in a phrase is a <word, stop-flag> pair. Hence, the format of the internal representation of the phrase is (<word, stop-flag> [<word, stop-flag>]). The Inverse Document Frequency (IDF) for each term is calculated and the key-terms in the index database are stored along with the IDF and document id. Each index is a 3-tuple of <key-term, IDF, doc-id>. These key-terms are processed by indexing module which rearranges the list of terms in location order as a list of terms in alphabetic order.

B. Sequential Pattern Mining Algorithm

Sequential Pattern mining is a data mining technique for finding statistically relevant patterns in the text data where the values are delivered in a sequence. SPMining algorithm can extracts the frequent sequential patterns from a document. Before starting the pattern mining process, data preprocessing is needed to improve the efficiency. Removing stopwords and term stemming algorithms are used. Feature selection is based on the term's TF*IDF value. Two parameters are needed for the method, PL (pattern length) and min_sup, the predefined minimum relative support to reduce the patterns with lower relative support discovered in a large document. Since the algorithm is a recursive function, the initial value of PL is the 1Term frequent patterns.

The document d = {S1, S2,…Sn}, where Si is a sequence represents a paragraph in d. P is a sequential pattern of d if there is a Si∈d such that P⊆Si. The *absolute support* of P, $supp_a(P) = |\{ S|S \in d \wedge P \subseteq S \}|$, is the number of occurrences of P in d. The *relative support* of P is the fraction of paragraphs that contain P in document d, denoted as $supp_r(P) = supp_a(P) / |d|$. A sequential pattern P is called *frequent sequential pattern* if suppr(P) is greater than or equal to a minimum support (min_sup). A frequent sequential pattern P1 is a *closed pattern* of P2, if P2 is a frequent sequential pattern, P1⊆P2, and suppa(P1) – suppa(P2) = 0.

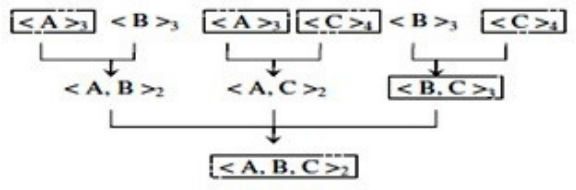

Fig. 2 Pattern Taxonomy of Closed Patterns

This algorithm includes the pruning phase to eliminate meaningless and redundant patterns. The Fig. 2 shows pattern <A, B> is a closed pattern of <A, B, C> i.e. they always appear in the same paragraph. Therefore, the shorter one (i.e. pattern <A, B>) is negligible and is considered as a meaningless pattern. Longer one is more meaningful and carries more information than shorter one. Therefore, after the pruning phase, only the significant patterns remain in the pattern taxonomy.

SPMining differs from other pattern mining algorithm because it can deal with several sequences at a time whereas others only handle one sequence at a time [15].

V. COMPARATIVE ANALYSIS

The challenging issue is to effectively deal with the large amount of text data. For this purpose various data mining techniques are developed based on the user's need what they want from text. The Rule-Based Approach for Phrase Extraction can extract the noun phrases from the text documents using name recognition and syntactic analysis modules. But the extracted phrases can have low frequency of occurrence as well as these phrases contain the redundant information. However in case of Sequential Pattern Mining algorithm it includes the pruning phase to eliminate the redundant and meaningless terms in pattern. In SPMining algorithm closed patterns are discovered, these patterns are more specific to the documents. For the text analysis purpose the pattern-based methods performs better than other retrieval techniques. The choice of text retrieval technique depends on what one regards as meaningful units of text and meaningful natural language rules for combination of these units.

VI. CONCLUSION

Text mining is the discovery of interesting knowledge in text documents. It is tedious task to find accurate knowledge from text documents. In this paper basic four methods for text representations are discussed to improve the effectiveness of the systems. The phrase based approach could not perform better than pattern-based approach. Many data mining techniques are developed for mining useful patterns. Sequential Pattern mining technique can improve the performance of system by using the closed patterns discovery and pruning phase. Using these technique main goals of text mining methods can be achieved by retrieving the information that meets users' needs within a relatively efficient time.

ACKNOWLEDGMENT

The author would like to thanks the university authorities and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing the infrastructure to carry out the work. This work is committed by university commission.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 1, January **2012**.

[2] Y. J. Fu., *Data mining: Tasks, techniques and applications,* IEEE Potentials, 16(4):18-20, **1997**.
[3] Sheng Tang Wu, "Knowledge Discovery using Pattern Taxonomy Model in Text Mining," Doctor of Philosophy Thesis, Queensland University of Technology, December **2007**.
[4] F. Sebastiani. "Machine learning in automated text categorization," ACM Computing Surveys, 34(1):1–47, **2002**.
[5] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management: International Journal, vol. 24, no. 5, pp. 513-523, **1988**.
[6] Yutaka Matsuo, Mitsuru Ishizuka "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information," FLAIRS **2003**.
[7] A. Inkeri Verkamo, Helena Ahonen-Myka, Oskari Heinonen, Mika Klemettinen, "Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery," proc. Of the workshop on Text Mining: Foundation, techniques and Applications, IJCAI, **1999**.
[8] H. Ahonen, 0. Heinonen, M. Klemettinen, and A. I. Verkamo, "Mining in the phrasal frontier," In Proceedings of PKDD, pages 343-350, **1997**.
[9] H. Ahonen, 0. Heinonen, M. Klemettinen, and A. I. Verkamo "Applying data mining techniques for descriptive phrase extraction in digital document collections," In Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries (ADL98), pages 2-11, **1998**.
[10] R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, **2003**.
[11] P D. Turney, "Learning algorithms for keyphrase extraction," Information Retrieval, 2(4):303-336, **2000**.
[12] Yongzheng Zhang, Nur ZincirHeywood, Evangelos Milios, "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora," WIDM ACM **2005**.
[13] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, **2006**.
[14] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, **2006**.
[15] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, **2007**.
[16] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, **2004**.

AUTHOR PROFILE

Vishakha D. Bhope doing M. Tech (Computer Science & Engineering) from Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431001, Maharashtra, India.
vishakhabhope@gmail.com

Assistant Professor Dr. Sachin N. Deshmukh Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431001, Maharashtra, India.
sndeshmukh@hotmail.com