ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017

Concept- Based Mining Model-A Sentence Based Concept Analysis

A Sentence Based Concept Analysis

S.Daisy Fathima Mary,
Lecturer, Department of computer science,
Thiruvalluvar University College of arts and science,
Thiruvennainallur
Daisyfatima_mca@yahoo.co.in.

Dr. C.Bhuvaneswari,
Assistant Professor & Head, Department of Computer Science,
Thiruvalluvar University College of arts and science,
Thiruvennainallur
Bhuvan.csdept@gmail.com.

***Abstract*— Text mining finds patterns and creates intelligence from unstructured text data. It is a statistical analysis of word frequencies within a document. Clustering is a technique for grouping the objects based on similarities. It finds the documents with common words and places the documents into cluster of common words. This paper describes a concept-based mining model which is used to analyze the term on each sentence by using semantic structure of each sentence and captures the tern frequencies. The similarity measures in the concept-based model are used to measure the importance of each concept with respect to the semantics of the sentence and the topic of the documents. The clustering quality is achieved more by the concept –based mining model in the sentence based analysis.**

***Index Terms* — Concept-based mining model, Clustering, Term frequency, semantic, similarity measures.**

I. INTRODUCTION

Text mining attempts to discover new and unknown information by applying techniques from natural language processing and data mining. It is important to note that understanding the meaning of words could not be deduced from statistical analysis of word frequencies. Natural language was developed for humans to communicate with one another and to record information, and computers are a long way from understanding natural language. Humans have the ability to understand the meaning of text and humans can easily overcome obstacles that computers cannot easily handle such as spelling variations and contextual meaning.

The concept mining creates the technology that combines the human way of understanding with the speed and accuracy of a computer. Concept mining is related to understand the meaning of text. Each word might have multiple meanings, and the context to disambiguate what is meant. To formalize the idea of meaning by linking meaning to concepts and multiple words might be used to represent a particular concept. Therefore, there is a need for a representation that captures the semantics in text in a formal structure.

II. LITERATURE REVIEW

"There is a ever-growing need to add structure in the form of semantic markup to huge amounts of unstructured text data. The technique of shallow semantic parsing, the process of assigning a simple WHO did WHAT to WHOM etc., structure to sentences in text. We formulate the semantic parsing problem as a classification problem using Support vector Machines." [1].

"An automatic clustering of nouns was performed using word co-occurrence data from a large corpus. This technique is based on the expectation that words with similar semantics will tend to co-occur with the same other sets of words. The clustering algorithm attempts to find such patterns of co-occurrence from the counts of grammatical relations between pairs of specific words in the corpus, without the use of any external knowledge or semantic representation." [2].

"The syntactic structure alone does not provide enough information for machine understanding of human language. The pragre tectogrammatics project endeavors to annotate semantic relationships at the same time as syntactic and morphological structure. The PropBank project at penn, which adds a layer of semantic annotation atop the syntactic structure."[3].

Clustering of web documents enables (semi) automated categorization, and facilitates certain types of search. Any clustering method has to embed the documents in a suitable similarity space. The four popular similarity measures (Euclidean, Cosine, Pearson Correlation and extended Jaccard) are compared in conjunction with several clustering techniques (random, self organizing feature map, hyper-graph partitioning, generalized K-Means, weighted graph Partitioning) on high dimensional sparse data representing." [4].

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017

"An unsupervised feature selection algorithm for data sets, large in both dimension and size. The method is based on measuring similarity between features whereby redundancy therein is removed. A feature similarity measure, called maximum information compression index, is introduced. The algorithm is generic in nature and has the capability of multiscale representation of data sets."[5].

III. TEXT MINING

Text mining refers to a collection of methods used to find patterns and create intelligence from unstructured text data. Text mining can be viewed as having two distinct phases such as term extraction and feature creation. Term extraction makes heavy use of string manipulation functions but also applies techniques from computational linguistics. Actual content is a result of the feature creation process. Feature creation applies unsupervised learning methods that reduce many potential features into a much smaller number of final variables. These features are then potentially useable as dependent or predictor variables in an analysis

Term extraction is the first step in deriving meaning or content from free form text. The next step is feature creation. The data can be represented as a rectangular array that has indicator variables for each term in the concept description. The terms are carefully analyzed and find some words denote similar and are unlike. Thus, the occurrence or non-occurrence of specific words may illustrate something useful about concept. One of the most common techniques used to group records with similar values on the terms together is known as cluster analysis.

A. Information Retrieval (IR)

Systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. The changes with the advent of digital libraries, where the documents being retrieved are digital versions of books and journals.IR systems allow narrowing down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, the mining information only about protein interactions, the analysis might restrict to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

B. Natural Language Processing (NLP)

NLP is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. For example:

- *Part-of-speech tagging* classifies words into categories such as noun, verb or adjective".
- *Word sense disambiguation* identifies the meaning of a word, given its usage, from among the multiple meanings that the word may have.
- Parsing performs a grammatical analysis of a sentence. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence.

C. Information Extraction (IE)

The process of automatically obtaining structured data from an unstructured natural language document. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include:

- *Term analysis,* which identifies the terms in a document, where a term may consist of one or more words. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers.
- *Named-entity recognition*, which identifies the names in a document, such as the names of people or organizations. Some systems are also able to recognize dates and expressions of time, quantities and associated units, percentages, and so on.
- *Fact extraction,* which identifies and extracts complex facts from documents. Such facts could be relationships between entities or events.

D. Different Levels Of Language Analysis

Knowledge about the structure of the language itself should be used in a natural language-system. This considerable knowledge refers to many aspects such as what the words are, what the words mean, how words join to construct a sentence, and how word meanings contribute to sentence meanings. The following are some different forms of knowledge that are relevant for natural language understanding.

- *Phonetic and Phonological Knowledge*: concerns how words are linked to the sounds that recognize them. This knowledge is central for speech-based systems.
- *Morphological Knowledge:* concerns how words are created from more basic meaning units called

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017

morphemes. A morpheme is the primitive unit of meaning in a language (for example, the meaning of the word "friendly" is derivable from the meaning of the noun "friend" and the suffix "-ly", which transforms a noun into an adjective).

- *Syntactic Knowledge:* concerns how words can be composed to construct correct sentence. It also concerns how to determine the structural role of each word in the sentence and what phrases are subparts of what other phrases.
- *Semantic Knowledge:* concerns what words mean and how these meanings can be combined in sentences to form the meanings of a sentence.
- *Pragmatic Knowledge:* concerns how sentences are used in different contexts and how context can affect the interpretation of the sentence.
- *Discourse Knowledge*: concerns the effect of the right away preceding sentences on the interpretation of the next sentence. This information is especially important for interpreting pronouns.
- *World Knowledge:* concerns the general knowledge about world structure that users of language should obtain. It also concerns what each language user knows about the beliefs and goals of other users.

E. Natural Language Text Processing Systems

A natural language text processing system may begin with morphological analyses. Stemming of terms, in both the queries and documents, is done in order to get the morphological variants of the words involved. Some NLP systems have been built to process texts using particular small sublanguages to reduce the size of the operations and the nature of the complexities.

F. The Role Labeling Task

There are many attempts to label thematic roles to a sentence automatically. There are two methods of semantic classification which are constituent-by-constituent (c-by-c) and word-by-word (w-by-w) classification. In the former, a deep syntactic parser is used to derive the constituents which are afterwards labeled with semantic roles. In the latter, the words are classified based on BIO tagging which has three cases: at the beginning of semantic role, inside semantic role, or outside semantic role. There are two main corpora: PropBank (UPenn) and FrameNet (UC Berkeley) that contain sentences tagged with semantic arguments. Other researches at Stanford University presented a model of natural language generation from semantics using the FrameNet semantic role and frame ontology. Some recent research work deals with the semantic role labeling as a chunking problem along with multi-class classification problem.

A semantic role labeler (or chunker) that groups syntactic chunks (i.e. base phrases) into the arguments of a predicate. This is accomplished by casting the semantic labeling as the classification of syntactic chunks (e.g. NP-chunk, PP-chunk) into one of several classes such as the beginning of an argument (B-ARG), inside an argument (I-ARG) and outside an argument (O). This amount to tagging syntactic chunks with semantic labels using the IOB representation. The chunker is realized using support vector machines as one versus all classifiers. It is important to note that extracting relations between verbs and between their arguments in the same sentence has a promising potential for understanding the meaning of a sentence.

IV. CONCEPT-BASED MINING MODEL

The concept-based model uses the semantic concepts extracted from documents and queries. Instead of matching the keyword features representing the documents and queries, the concept-based techniques attempt to compare the semantic concepts of documents to those of given queries. The similarity of documents to queries is determined by the matching level of their semantic concepts.

A raw text document is the input to the concept-based mining model. Each document has well defined sentence boundaries. Each sentence in the document is labeled automatically based on the PropBank notations. After running the semantic role labeler, each sentence in the document might have one or more labeled verb-argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labeled verb-argument structures includes many verbs associated with their arguments. The labeled verb-argument structures, the output of the role labeling task, are captured and analyzed by the concept-based model on the sentence, document, and corpus levels.

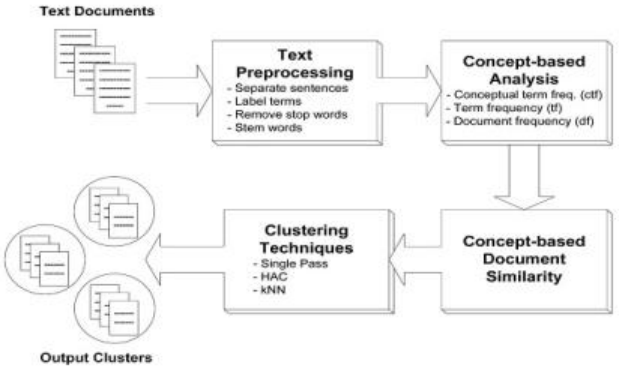
In the concept-based model, both the verb and the argument are considered terms. One term can be an argument to more than one verb in the same sentence. This means that the term can have more than one semantic role in the same sentence. In such cases, the term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based model, a labeled terms either word or phrase is considered a concept.

V. CONCEPT-BASED STATISTICAL ANALYZER

The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017



- *Sentence-Based Concept Analysis:* The sentence-based concept analysis is used to analyze each concept at the sentence level a new concept-based frequency measure is used called the conceptual term frequency (ctf). The ctf calculates the number of occurrences of concepts in the sentence.

When the sentence is given as the input, the sentence based concept analysis finds the total number of sentences, and the total number of sentences containing the concept. The ctf calculations of concept c in sentence s and document d are as follows:

a) Calculating ctf of concept c in sentence s: The calculation of ctf in sentence level finds the concept c, which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of s and also the ctf finds the number of occurrences of concept c in verb argument structures of sentence s. The calculation of the ctf is a local measure on the sentence level.

b) Calculating ctf of concept c in document d: For the calculation of ctf in the document level, the sentences in the documents are found based on the termination symbol (.) and all the sentences are listed in the array structure for the calculation of ctf. Then it finds the concepts in each sentences and calculates the ctf by incrementing its term frequency count.

In a document, concept c can have many ctf values in different sentences in the same document d. Thus, the ctf value of concept c in document d is calculated by*For Authors of More than Two Affiliations:* To change the default, adjust the template as follows.

$$Ctf= ctfn/sn$$

Where sn is the total number of sentences that contain concept c in document d. Taking the average of the ctf values of concept c in its sentences of document d measures the overall importance of concept c to the meaning of its sentences in document d. A concept, which has ctf values in most of the sentences in a document, has a major contribution to the meaning of its sentences that leads to discover the topic of the document. Thus, calculating the average of the ctf values measures the overall importance of each concept to the semantics of a document through the sentences..

EXAMPLE: The calculation of ctf in a document.

Consider in a document a concept c which appears twice in document d in the first and the second sentences. The concept c appears five times in the verb argument structures of the first sentence s1 and three times in the verb argument structures of second sentence s2. In this case the ctf value of concept c is equal to = 4

- *A Concept-Based Similarity Measure:* A concept-based similarity measure, based on matching concepts at the sentence, document, corpus and combined approach rather than on individual terms (words) only, is devised. The concept-based similarity measure relies on three critical aspects. First, the analyzed labeled terms are the concepts that capture the semantic structure of each sentence. Secondly, the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence, as well as to the main topics of the document. Finally, the number of documents that contains the analyzed concepts is used to discriminate among documents in calculating the similarity. These aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the sentence-level by the ctf measure, document-level by the tf measure and corpus-level by the df measure. The concept-based measure exploits the information extracted from the concept-based analysis algorithm to better judge the similarity between the documents.

This similarity measure is a function of the following factors:

1. The number of matching concepts, m, in the verb arguments structures in each document (d),

2. The total number of sentences, s, in each document d,

3. The total number of the labeled verb argument structures, v, in each sentence s,

4. The ctfi of each concept ci in s for each document d where (i = 1, 2,…,m)

5. The tfi of each concept ci in each document d where (i = 1, 2,…,m),

6. The dfi of each concept ci where (i = 1, 2,…,m),

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017

7. The length, l, of each concept in the verb argument structure in each document, d, and

8. The length, Lv, of each verb argument structure which contains a matched concept.

9. The total number of documents, N, in the corpus.

The conceptual term frequency (ctf) is an important factor in calculating the concept-based similarity measure between documents. The more frequent the concept appears in the verb argument structures of a sentence in a document, the more conceptually similar the documents are. The concept-based matching consists of either an exact match or partial match between two concepts. Exact match means that both concepts have the same words. Partial match means that one concept includes all the words that appear in the other concept.

VI. CONCLUSION

Most of recent text mining systems consider only the presence or the absence of keywords in text. Statistical analysis of word frequencies is not sufficient for representing the meaning of text. Concept-based mining is targeting the semantics of text rather than word frequencies.

A concept-based model which captures and represents the semantics in text based on concepts. The concept-based model discovers the structured knowledge to be utilized in several applications. The new concept-based model is proposed to improve the text clustering qualities. By exploiting the semantic structure of the sentences in documents, a better text clustering results is achieved.

REFERENCES

[1] S. Pradhan, K. Hacioglu, W. Ward, J.H. Martin, and D. Jurafsky, "Semantic Role Parsing: Adding Semantic Structure to Unstructured Text," Proc. Third IEEE Int'l Conf. Data Mining (ICDM), pp. 629-632, 2003

[2]. D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Roles," Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002

[3]. P. Kingsbury and M. Palmer, "Propbank: The Next Level of Treebank," Proc. Workshop Treebanks and Lexical Theories, 2003

[4] M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, July 1980.

[5]. A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf. Artificial Intelligence: Workshop Artificial Intelligence for Web Search (AAAI), pp. 58-64, 2000.