A Survey Paper on Concept Mining in Text Documents

K. N. S. S. V. Prasad
CSE Dept.
MANIT Bhopal 462003, MP

S. K. Saritha, PhD
CSE Dept.
MANIT Bhopal 462003, MP

Dixa Saxena
CSE Dept.
MANIT Bhopal 462003, MP

ABSTRACT

Concept Mining has become an important research area. Concept Mining is used to search or extract the concepts embedded in the text document. Concept based approach search for the informative terms based on their meaning rather than on the presence of the keyword in the text.

Keywords

Concept mining, Term Frequency, Inverse Document Frequency, Conceptual Term Frequency

1. INTRODUCTION

Data Mining is about finding interesting and useful patterns from data. Mining can be done in text, images, videos, and so on. Text Mining **[1]** is one of the data mining techniques that attempts to discover new, previously unknown information by applying techniques from Natural Language Processing. Text mining is different from what are familiar with in web search. Mostly user looks into already existing data which is being written by others. The problem is pushing aside all the material that currently is not relevant to your needs in order to discover the relevant information. Text mining has different names i.e., Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) **[3],** generally defined as the process of extracting interesting and non-trivial information and knowledge from non-structured text. Text mining is a young interdisciplinary field which appeals on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information is stored in the form of text, text mining is considered to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The task of Knowledge Discovery from Text (KDT) **[3]** is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) methods. Its aim is to get insights into large quantities of text data. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining **[1]** is alike to data mining, but In data mining the tools**[2]** are constructed to handle structured data from databases ,while text mining can work with unstructured or semi-structured data sets For example emails, full-text Documents, html files etc. As a result, text mining is a much better solution for companies.

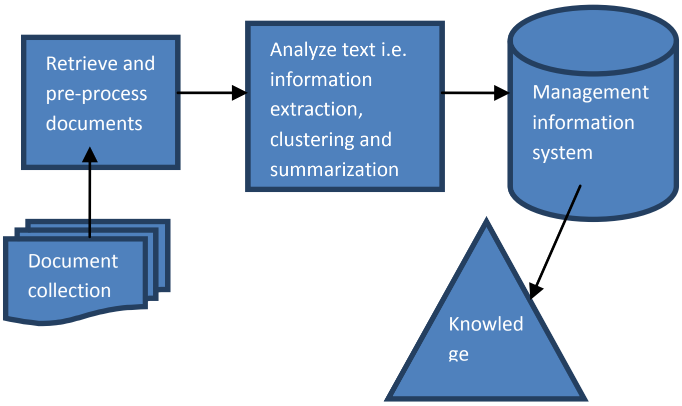
Retrieve and pre-process documents
Analyze text i.e. information extraction, clustering and summarization
Management information system
Document collection
Knowled ge

Fig 1. An Example of Text Mining

1.1 Concept Mining [4] is used to search or extract the concepts embedded in the text document. Concept based approach search for the informative terms based on their meaning rather than on the presence of the keyword in the text. Concepts are terms or set of terms with some meaning or relation between terms. Sometimes few terms came together very frequently in a text that relatedness in terms also lead to a concept. The meaning of concept sometimes very different from original meaning of words occurred in the concept.

Ex. - White house.

A concept may carry different importance in different sentences/ document/corpora.

1.2 Techniques used in Concept Mining:

1.2.1 Term frequency is often used as a weighting factor in text mining, depends on the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

1.2.2 Term Frequency Inverse Document Frequency (TFIDF): The TFIDF or Term Frequency Inverse Document Frequency method measures the significance of a term in a document within a set of documents.
TFIDF calculated as:

$$TFIDF\ (ti, dj) = TF(ti, dj)IDF(ti)$$

Where TF (ti, dj) stands for Term Frequency and is defined by

$$TF\left(ti,dj\right)=\begin{cases} N\left(\frac{\left(ti,dj\right)}{\left|d\right|}\right) & if\ N\left(ti,dj\right)>0 \\ 0 & Otherwise \end{cases}$$

Where ti is a term of document dj, $N(ti,dj)$ denotes the frequency ti in dj, and $|dj|$ is the total number of tokens in document dj. $IDF\ (ti)$ Stands for Inverse Document Frequency and is defined as

$$IDF(ti)=\log\left(\frac{Tr}{df\,(ti)}\right)$$

Where $df\,(ti)$ stands for Document Frequency of term ti and denotes the number of document Tr in which ti occurs at least once. Terms that occur in a large number of documents tend to be stop words. By using TF × IDF **[6], [7]** calculation, it is expected that stop words can be discriminated by their TFIDF score. Based on the characteristics of the logarithmic function of $\frac{Tr}{df\,(ti)}$, stop words tend to have TFIDF scores close to or Equal to zero. Keywords are considered important if they have high TFIDF value and high Document Frequency (DF) value. We say that DF value is high if it is greater than a given threshold value. Typically, the TFIDF value is an indicator to identify keywords, and the DF value is an indicator to identify interesting keywords.

1.2.3 Conceptual Term Frequency (CTF):

The ctf is calculated on both sentence level and document level. The ctf can be defined as number of occurrences of concept in sentence/document level.

1.3 Problems Faced in Concept Mining

The mappings of words to concepts are often ambiguous. Each word in a given language will relate to several possible concepts.

1.4 Advantages of Concept Mining

Text mining models are very large when compared to concept mining. The model that is going to classify, for sample, news stories using Support Vector Machines or the Naïve Bayes algorithm will be very large, in the megabytes (mb), and it takes much time to load and evaluate. Concept mining models can take very short time for comparison - hundreds of bytes. For some applications, such as similar meaning detection, concept mining offers new possibilities. Where the writer has been cunning enough to perform a thesaurus based substitution that will fool text comparison algorithms, and the concepts in a document will be relatively unchanged.
Ex-**‘The cat sat on the mat**” is meaningfully same as **‘The feline squatted on the rug’**

Concept Mining Applications

1.) Detecting and indexing similar documents in large corpora.
2.) Clustering.
3.) Classification.
4.) Natural language translation.

2. LITERATURE REVIEW

2.1 In [11] the main objective is to perform clustering and concept mining. The author proposed a concept-based mining model. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, Corpus-based concept-analysis and concept-based similarity measure.

In concept based mining model, author tried to say that, it takes raw text document as input and gives clusters as an output

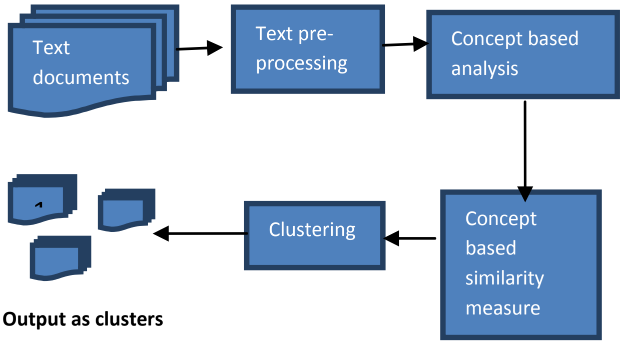

Fig.2.Concept-based mining model system.

In concept based term analysis, author tried to say that, the main aim of this task is measure concept based term analysis on sentence and document levels. Author used CTF (Conceptual Term Frequency) for analysing concepts at document/sentence level

In concept based similarity measure the author tried to say that how to calculate similarity between documents**.** The similarity between two documents can be calculated as:

$$similarity(d1,d2)=\sum_{i=1}^{m}\max\left(\frac{li}{si1},\frac{li}{si2}\right)*weighti1*weighti2$$

Where

$$weighti1=tfweighti1+ctfweighti1$$
$$weighti2=tfweighti2+ctfweighti2$$

After concept based similarity measure the concepts in a document are send for clustering. For clustering some techniques are used. They are HAC, SINGLE PASS CLUSTERING, and KNN. And these techniques give an output as set clusters as shown in the **figure 2**.

As compared to traditional analysis of a term (word or phrase), the concept based mining results gives fast and better results substantially

2.2 In [12] the main objective is to Build/Enrich WorldNet Dictionary and clustering. In Natural Language Processing Word Net dictionaries are frequently used. Most of the paper uses these dictionaries; but still the enrichment of Word Net dictionaries needs concern. Improvement in this library will definitely help researchers for getting better results. The techniques generally used to improve them are follows automotive strategies of converting text from a source language to another destination language. Most of these techniques follow blindly an unspoken rule: get predefined rules of the language and use it in as dictionary or a simpler parser having grammatical rules. ALOC approach uses to find similar conceptual ideas, and can be useful in Word Net dictionary building as these dictionaries are also stores items by their conceptual meaning.

So the ALOC approach is used in this paper, based on conceptual meaning and distance of conceptual same meaning words but described approach was on English language only. It can be extended to other language also.

Limitation of this paper is the approach was presented on a very restricted domain of samples, which were all in English. Furthermore research should focus on extending the domain of inputs and languages

2.3 In **[13]** the main objective is to perform concept mining and clustering. The approach consists of

a) TFIDF

The TFIDF or Term Frequency Inverse Document Frequency method measures the significance of a term in a document within a set of documents.

b) Co-occurring Keywords

A set of one keyword describes a concept. However, a set of more than one keyword might describe a new meaning that is beyond the meaning of each individual keyword. If this new meaning is important, the set of these keywords will consistently appear in the form of co-occurring keywords. Concepts within documentscan be captured through high frequency and co-occurring keywords.

c) Association Rule for Mining Keyword sets

Association Rule Mining **[8]** is used to show relationships between keywords. Many algorithms have been developed to find association rules. The most popular algorithm is The Apriori Algorithm. The algorithm generates n-keyword sets from n− 1- keyword sets, where n > 1. This means that the process of generating n-keyword sets depends on the previous step when generating n − 1-keywordsets.

Concept in a text document is represented by graph called simplicial complex where vertex and edges are keyword and relation between the keywords. This structure named simplex. A simplex with high frequency shows that relation between that keyword is more frequently encountered than others in same text. This relation contains a concept. This simplex relation is figured using Association rule mining. All the simplices of the text collectively show the concept structure of the text.

A text may contain a number of concepts. This concept may appear in more than one class or cluster so it’s not possible to have a separate line between two or more text. In the paper it is described that similar meaning text may have same or nearly same graphical structure of concepts. We can also compare graphical structure of different language to check whether associated text is meaning wise same or not.

Limitation of this paper is**,** each document may have several concepts. Therefore, many documents may be intertwined among themselves. Consequently, and cleanly separating documents may not always be possible.

2.4 In [14] the main objective of this paper is web page clustering and concept mining

The author described three important terms that play important role in webpage clustering. They are TF-based, TF-IDF based, TF+filtering. The TF-based, TF-IDF based are explained above **[6][7].**

TF + filtering: Extract terms with high frequency but filter those words, which appear often on web pages without any special meaning, such as download, link, news and so on.
The author summarised webpage clustering with four steps
1) Term extraction (i.e. Feature selection)
2) Term-Document matrix generation
3) SVD –based clustering
4) Merge results if necessary.

Based on clustered results, concept mining consists of three steps: first mining the word sets that frequently appear together on web pages based on Apriori Algorithm [9] then, Generatingconcepts item sets using the algorithm, finally building a concept hierarchy based on theiroverlap of itemsets. Here, concept item set is defined as co-occurring word sets like **[10].**

Limitation of this paper is the approach was presented on a very restricted domain of examples, which were all in English and Chinese. Further research should focus on extending the domain of inputs and languages.

3. CONCLUSION

At last we conclude that, in text mining, Concept based clustering targets the meaning of words/sentence. It has given significant improvement over traditional term frequency. Concept mining calculates the contribution of words to the meaning of the sentence, which implies a more efficient and sensible clustering. Concept may be a word or set of word which gives meaningful contribution to the text but we can found other concept in same or different document, which gives same or nearly same meaning. This meaning wise same concept must be treated as single entity while counting contribution to text. In this approach, same meaning concept is grouped together, called set of concept. Set of concept can be seen as same meaning but different word tokens. The clustering will be done based meaning of set of concept. A natural language text contains various words, some of the text may give higher contribution to text meaning than other words. Sometimes combination of words contributes more than the individual words. Extracting text entities which describe meaning of text in called concept forming and these extracted entities are called concepts. Concepts may be words or a meaningful set of words which comes together in text more frequently.

3.1 Concept Extraction Steps

1. Pre-processing of text.

2. Finding Verb-Argument structures.

3. Argument reduction.

4. Verb-Arguments frequency Count.

5. These Collected terms are now Concept.

3.1.1 Pre-processing of text

The row text may be unstructured and must be processed to tabulated form. The common steps to pre-processing row text is-

a.) **Tokenizing** - Text tokenizing is a process of fragmenting of text into words. The result is a list of independent words called token.

b.) **Stop Words removal** - The list of tokens may contain stop words which usually doesn't give any information about the text. So removal of stop words takes place after tokenizing.

c.) **Word stemming** - stemming is the process of ruling back a word to its root form from its derived form. It ensures that all the tokens are in its root form.

3.1.2 Finding Verb-Argument structure

Verb describes as action in a sentence argument usually helps the verb to define the meaning of sentence so a verb-argument pair is closer to meaning of sentence than randomly taking word tokens.

3.1.3 Argument reduction

An argument can be associated with more than one verbs this results redundancy in arguments. Similar arguments which gives same meaning or nearly same meaning with various verbs in a sentence is reduced to one joins the verb which have higher argument count.

3.1.4 Verb-Argument frequency count

Verb-argument with high frequency tells this pair of word argument is giving more contribution to sentence meaning so verb-argument with low frequency is less concerned. Frequency count is calculated of all verb-argument tokens and low frequency count tokens are trimmed out.

4. REFERENCES

[1] Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43

[2] Navathe, Shamkant B., and Elmasri Ramez, (2000), "Data Warehousing and Data Mining", in "Fundamentals of Database Systems", Pearson Education pvtInc, singapore, 841-872.

[3] HaralamposKaranikas and BabisTheodoulidis Manchester, (2001), "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK

[4] https://en.wikipedia.org/wiki/Concept_mining

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

[6] G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol5, 513-52

[7] G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol 5, 513-523

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, 1994

[9] Agrawal R, Imielinski T, Swami A, "Mining association rules between sets of items in large databases". Proc of the 1993ACM SIGMODInternational Conference on Management of data

[10] Bing Liu, Yiming Ma, "Discovering unexpected information from your competitors 'Web Sites in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, USA.

[11] An Efficient Concept-Based Mining Model for Enhancing Text Clustering, Shady Shehata, Member, IEEE, FakhriKarray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE 2010

[12] Concept mining from natural language texts, Rockai V. Dept. of Cyber. & Artificial Intelligent, Tech. Univ. of Kosice, Kosice, Slovakia Mach. M IEEE 2012

[13] Concept Mining using Association Rules and Combinatorial Topology Sutojo, A, San Jose State University, San Jose IEEE 2007

[14] Webpage Clustering and Concept Mining, an Approach to Intelligent Information Retrieval. Fang Li, Martin Mehlitz, Li Feng, Huanye Sheng, DEPT of CSE, Shanghai Jiaotong University, Shanghai ,China IEEE 2006