Applied GIS

ISSN: 1832-5505
Vol-5 Issue-02 April 2017

Review of Models for Concept-Based Mining to Improve Text Clustering

Palvi Sharma[#1], Sonika Gupta[#2]

[#]*School of Computer Science& Engineering, Shri Mata Vaishno Devi University, Kakryal, Katra, Jammu & Kashmir, India*

Abstract- The internet is home to an enormous trove of information in the shape of articles, studies, weblogs, emails, messages, and more. Data mining is the process of extracting useful information from large amounts of unstructured data sets, such as those found on the internet. Extensive study is done to advance the methods that may improve text mining. It is the statistical analysis of the "term," which may be a word or a phrase, that forms the basis of the most common text mining techniques. The importance of a "term" is only recorded between documents in a term frequency (tf) analysis. The "term" is analyzed at the sentence, document, and corpus levels in the present methodology, which is termed the Concept-based mining model. Each "term" in a phrase is assigned a semantic function as a concept in the Concept-based mining approach. The concept recognition process in NLP relies heavily on the sentence structure provided by the verb argument. This study analyzes the approaches used in the field of Text mining and reviews the existing literature on the topic. The gap between text mining and natural language processing methods is bridged with a focus on concept-based mining methods.

Keywords- **Clustering Methods, Mining Models Based on Concepts, and Text Mining.**

INTRODUCTION

E-mails, papers created in other word processors, movies, photographs, audio files, presentations, websites, and many other types of business documents are just some of the unstructured data that flood the internet every day. The amount of unstructured data available online is increasing quickly. It's expanding at a far quicker rate than organized data online.[1] Professionals have determined that eighty percent to ninety percent of a company's data is unstructured. Because the amount of data being saved is growing exponentially every day, [2] there is a pressing need to mine the unstructured data into meaningful information so that valuable insights may be drawn from the ever-increasing volume of facts. It's a crucial component of NLP (Natural Language Processing). [3]With text mining, enormous amounts of unstructured data are inspected and analyzed. Unstructured data is mined for previously unseen insights. Data is mined from textual materials using text mining. Term Frequency (tf) is a metric used to establish a word's significance in a text. There is a possibility that two or more than two words have the same frequency in the text, but only one phrase with the same frequency might contribute to the overall explanation of the whole sentence [4]. That significant phrase has to be grouped differently from the other terms that follow having the same frequency. There is a potential for examining words inside a sentence because of the connection between the verb and their argument structure. Each word has a specific meaning in the context of the phrase, and this meaning contributes to the overall meaning of the sentence. [5]The best method for imparting structure to the phrase is the use of verb argument structure. The sentence is given a label using a semantic role labeler. The notions that each word plays in a phrase are its meaning. A concept may be an additional word or phrase in a sentence, where each word or phrase lends meaning to the sentence's theme. Data mining methods center on the document's underlying ideas and phrases. The current methods of data mining rely on the calculation of the frequency of terms inside a given text. When searching for a phrase among a large collection of documents, it might be challenging to isolate those that include that term without also taking into account whether or not the "term" inside those documents really contributes to the document's meaning. Therefore, it is important to learn the terminology used to determine the subject matter of a document and capture the meaning of its words. Sentence meaning is analyzed by semantics. [8] By scanning the newly added document and extracting the matching ideas, the mining model is able to detect a concept match from this document to all previously processed document in the data set.

Few of the important terms used in the paper are:

- Verb-argument structure: e.g. Palak plays guitar. "**Plays**" is the verb. "**Palak**" and "**guitar**" are the arguments of the verb "**plays**".
- Label: A label is assigned to an argument, e.g. "**Palak"** has subject (or agent) label. The "**guitar**" has object (or theme)label.
- Term: Term can be an argument or verb.
- Concept: Concept can be recognized by using NLP. Each term which possesses semantic role within the sentence is called concept.

ISSN: 1832-5505
Vol-5 Issue-02 April 2017

Clustering, an unsupervised learning approach, is the most significant data mining methodology [3]. Clustering is a method for organizing large datasets into manageable groups of similar records. Group items that share characteristics into distinct categories within the collection.Clustering is preferable to classification because it may be tweaked to better isolate the characteristics that distinguish across groupings. [4] Decision trees, rule-based systems, K-means clustering, single-pass clustering, neural networks, clustering based on data summarization, and hierarchical clustering are only a few of the approaches utilized for text clustering. The Vector Space Model is a popular method for grouping and classifying text documents. The model might either be in two or three dimensions. The VSM is an example of a model that relies on textual similarities. Similarity between the document and the query is shown by this. Each "term" represents a separate dimension in the multidimensional space in which documents are represented. The document's Term might be any random word or phrase. In a vector space model, a document's features are represented by a vector. A feature vector calculates the frequency and similarity of the phrase in a text. There are several different similarity metrics that may be used to detect commonalities in papers. Some examples of similarity measures are the Cosine Similarity and the Jaccard Similarity.
Cosine similarity is a measure of how closely two papers are alike by determining the cosine angle between them.
The Jaccard index compares two strings based on the proportion of common phrases to total terms.

• Both TF and IDF are employed as primary characteristics in text mining. It's a measure of how important a word is in a certain context (document, phrase, or corpus). In the fields of information retrieval and text mining, this is a type of weighting. The value of TF and IDF increases in proportion to the frequency with which a word occurs in the manuscript. The effectiveness of a clustering algorithm may be improved by appropriately weighting each component.

While the prior mining model did a good job of capturing sentence meaning, it was not able to accurately determine how often a given phrase appeared in the text. The concept-based mining technique provides an in-depth analysis of ideas at three distinct levels: the phrase, the text, and the corpus.

CONCEPT BASED MINING MODEL

Sentence-based concept analysis, document-based idea analysis, corpus-based concept analysis, and a concept-based similarity measure are the four components of the novel Concept Based Mining Model proposed here, which examines each concept and enhances the text clustering standard. Each idea is based on the syntax and semantics of a single phrase. Even if they both appear often in the same phrase, one word is much more significant than the other. The concept-based mining model receives its input from a raw text data set.An Initial Processing of Text
This is the initial step in the concept-based mining methodology, and it is very important to the field of text mining. The text pre-processing step is fed with raw text data. This study focuses on four procedures that are Words, terms, and stop words are separated, labeled, and removed, and stemmed.First, you need to extract the text from the document. There is a distinct dividing line between sentences in every paper.Each phrase in a document has been assigned a label by a Semantic Role Labeler. Each phrase in the text includes many different verb argument forms when using semantic role labeler. Sentence meaning is organized by the semantic structure of the words used. A tagged term is a word or phrase that is picked up on when reading a text in the Concept Based Mining Model. Because of its flexible verb argument form, the phrase may have numerous purposes within the context of the sentence. The results of the labeling job at the sentence, document, and corpus levels are captured and analyzed using a concept-based mining model.

In natural language processing, "stop words" are the common terms that should be eliminated. Before the text is preprocessed, stop words are eliminated. Stop words like "a," "of," "the," "which," "who," "is," "at," "in," and "on" are present in this world. Problems with phrase searches are often caused by stop words. Prop Bank Notations will eliminate any and all stop words from a manuscript. By omitting these filler words, readers are free to concentrate on the text's actual contents. Stop words are introduced in natural language processing to eliminate useless words (information).

A list of common prefixes and suffixes that might be found in an inflected word is used by stemming algorithms to remove the beginning or end of the term. Porter's algorithm for stemming English text has been created. There are now less of the document's unnecessary words.

ISSN: 1832-5505
Vol-5 Issue-02 April 2017

TABLE 1
STEMMING EXAMPLES

Form	Suffix	Stem
Studying	-ing	Studi
Closely	-ly	Close
Played	-ed	Play
Affected	-ed	Affect
Amusing	-ing	Amus
Gratefully	-fully	Grate

B. Concept Based Analysis

In the idea based mining paradigm, this is now Step 2. Once the text has been pre-processed, it is sent into the Concept Based Analysis step. Every notion is broken down into sentences and analyzed using Sentence Based notion Analysis. Analyzing each idea in each document is what Document Based idea Analysis does. Each and every notion is examined using corpus-based concept analysis. When compared to single-term analysis, concept-based analysis aims to correctly analyze concepts at the sentence, document, and corpus levels.

• phrase-based concept analysis (ctf): a novel concept-based frequency measure, the conceptual term frequency (ctf), is generated to analyze each concept at the phrase level. Ctf measures how many times the notion c occurs in the phrase s.

• Document-Based Concept Analysis (tf): This method uses the concept-based term frequency tf and the number of occurrences of a concept in a document to analyze each concept at the document level.

Concept Analysis Based on Corpora (df) Differentiating ideas are extracted, and the number of documents that include a given concept (c) is determined using a concept-based document frequency (df) measure. The study of language makes use of corpora, which are enormous collections of texts.

C. **A Similarity Metric Based on Concepts**

This similarity metric prioritizes matching concepts over individual phrases, and so operates best at the sentence, document, and corpus levels. There are three primary features of a concept-based similarity measure. First, the analyzed label terms capture the sentence's semantic structure. Second, the frequency of each idea is utilized to evaluate its significance in establishing the sentence's meaning and developing the essay's central argument. Third, while determining how similar two papers are, the number of occurrences of the topics under analysis is employed as a differentiating factor. A conceptual term frequency (ctf) is a crucial element in determining how similar two papers are to one another. In concept-based matching, ideas may either be a perfect match or a partial match. When two ideas are a perfect fit, their corresponding words are also identical. [3] Jaccard's distance and proximity measures are used to determine how similar the papers are to one another.

How Far Is the Jaccard Distance? The Jaccard distance between two items illustrates how unlike they are to one another.

A measure of proximity reveals the degree of resemblance between two entities.

Cluster Analysis Methods

Data mining's process is known as text clustering. Text clustering, also known as document clustering, is a powerful method for organizing texts. Clustering is the process of grouping collections of items such that those in the same group have more similarities than differences. Text documents in a collection are clustered. The effect of concept-based similarity on clustering is evaluated using a small set of popular document clustering methods.

Clustering, Hierarchical:

To do hierarchical clustering, [5] clusters must be created with a certain, ordered hierarchy. Hierarchical clustering may go one of two ways: either dividing the data into smaller groups or lumping them together.

Bottom-up agglomerative clustering begins with each data point in its own cluster. Each time, we take the two nearest clusters and merge them into one.

Top-down, divisive clustering places all data points in a single cluster at the outset, then divides them into subclusters through recursive partitioning.

Hierarchical agglomerative clustering essentially boils down to repeatedly merging two neighboring clusters inside a larger cluster. The average of the points forms the centroid, which we may use to symbolize the cluster.

1) K means:

K means is relatively an efficient method. K means is an iterative algorithm that attempts to separate the data sets into K pre- defined separate disjoint piece of clusters where each data is part of only one cluster. It attempts to generate the Inter-Cluster data points as likely while also keeping the clusters as far as possible. It provides data points to a cluster, in such a way that the addition of the squared distance within the data points and the clusters centroid is at the least possible. The small difference in some way that is within the clusters, the more identical data points are between the same clusters. K-means and hierarchical

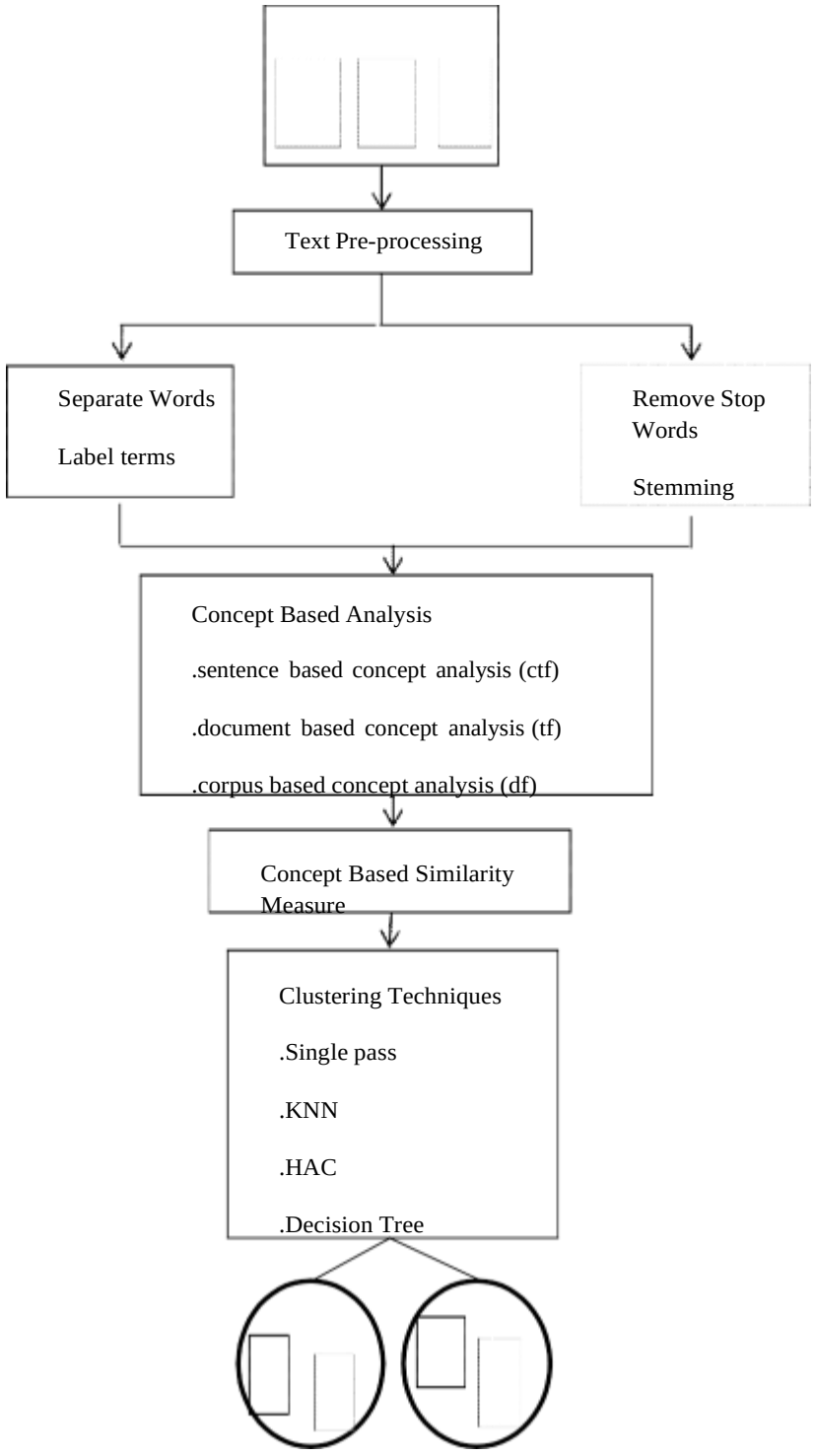

Fig. 1 Concept Based Mining Model

LITERATURE SURVEY

The goal of this literature survey to analyse the existing document clustering system and their advantages and disadvantages. We have analysed that [4] S. Shehata, F. Karray, and M. Kamel, presents a system based on semantic analysis at sentence and document level and it was not able to find similar document at corpus level

[5] PRADNYA S.RANDIVE & NITIN N.PISE examines every concept at the sentence, document and corpus level that increase similarity measure based on the term analysing model of document. The results are calculated by using F-measure and Entropy.

[6] S.Brindha, Dr. K.Prabha, Dr. S.Sukumaran analyses concept at the sentence and document level. They enhance the supervised weighting in the TFIDF model. They examine distinct weighting scheme, makes use of a type of information ratio to decide a term's contribution for category ahead with class information. They analysed the term weight based technique to analyse which isbelonging to text classification technique as well as text document technique.

[7] M. K. Vijaymeena and K. Kavitha talk about the three similarity approaches such as String-similarity measure, Corpus based similarity and Knowledge based similarity.

[8] T. Svadas and J. Jha present an ontology which could be consider as a repository of knowledge in which concepts and terms are define and also finds the relationship between these terms and concepts. Ontology makes the knowledge that is implicit for humans and explicit for computers. This can increase the text mining for a particular reason.

TABLE 2
COMPARISON TABLE

S.NO	Author Name	Year of Publicati on	Data Set Used	Similarit y measure used	Clusteri ng Techniq ue
1	Shady Shehata, Mohame d S.Kamel, Fakhri Karray	2010	ACM abstract articles, Documen ts from Reuters, Samples from Brown Corpus, Messages collected from Usenet groups	F-measure, Entropy	Hierarchi cal Agglome rative Clusterin g, Single pass, K-Nearest Neighbo ur
2	Pradnya Randive ,Nitin Pise	2012	Web Documen t	F-measure Entropy	Single pass, HAC algorithm KNN algorithm
3	Gausiya Begum, N.Murrat sultana	2013	Text Documen t	Cosine measure, Jaccard Distance, Proximit y measure	HAC KNN
4	Ms Trupti U. A hirrao, Dr. Varsha H.Patil	2014	Web documen t	F-measure, Entropy	Partitiona l clusterin g, Hierarchi cal Clusterin g
5	Gulmoha med Rashita Banu	2015	ACM, Reuters, Brown Corpus, Usenet News groups	Cosine measure, Jaccard measure	HAC, KNN, Single pass, Multipas s
6	Twinkle Svadas, Jasmin Jha	2015	20 News groups	Precision , Recall, F-measure	HAC, Single pass
7	S.Brindh a, Dr. K.Prabha , Dr. S.Sukum aran	2016	Reuters, ModApte	Cosine Similarit y, Precision , Recall	KNN, Decision tree
8	PRADN YA S.RAND IVE, NITIN N.PISE	2017	Web Documen t	Precision , Recall, F-score	Single pass, HAC,K-Nearest Neighbo ur

CONCLUSIONS

It's a link in the chain between NLP and text mining. A novel concept-based mining approach is created, and its four components improve text clustering quality. The concept-based mining model consists of the following four parts: Sentence-level (ctf) concept analysis examines each idea individually. Document-based concept analysis (tf) examines individual concepts within the context of each individual document. Analysis of each topic in a corpus, depending on how often it appears in documents; this metric is called document frequency (df). When comparing documents in a corpus, a concept-based similarity measure takes into account the importance of every concept related to the sentence's meaning, the document's subject, and the documents' distance from one another.

REFERENCES

[1] "Unstructured data mining and its applications," Int. J. Curr. Eng. Sci. Res., vol. 3, no. 3, pp. 36-40, 2016; J. J. Wagh, J. D. Gondane, and A. T. Dukare.

[2] (Reference) [2] O. Rusu et al., "Converting unstructured and semi-structured data into knowledge," Proc. - RoEduNet IEEE Int. Conf., no. January, 2013.

[3] The article in question is "a Concept-Based Mining Model for Increasing Text Performance," which appeared in 2011's volume four, issue four, pages 398–400.

[4] "An efficient concept-based mining model for enhancing text clustering," by S. Shehata, F. Karray, and M. Kamel, IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pages 1360-1371, 2010.

[5] Reference: [5] A. Nirmala and K. A. Vanitha, "Effective Concept-Based Mining Model For Text Clustering," Volume 4, Issue 7 (September 2013), Pages 1473-1480.

[6] "Concept Based Mining in Text Clustering," by P. Randive and N. Pise, pp. 81-84.

[7] [7] "the Comparison of Term Based Methods Using Text Mining," S. Brindha, K. Prabha, and S. Sukumaran, vol. 5, no. 9, pp. 112-116, 2016.

[8] For example: [8] M. K. Vijaymeena and K. Kavitha, "a S Urvey on S Imilarity M Easures in T Ext M Ining," Mach. Learn. Appl., an International Journal, volume 3, issue 1, pages 19–28, 2016.

[9] As cited in [9] by T. Svadas and J. Jha in "Document Cluster Mining on Text Documents," volume 4, issue 6, pages 778-782, 2015.

[10] "A study on social big data analysis using text clustering," by J. H. Ku and Y. S. Jeong, vol. 7, no. 2, 2018, pp. 1–4.

[11] As cited in [11] "Enhancing text clustering using concept-based mining model," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 1043-1048, 2006.

[12] According to [12] L. Zahrotun's "Comparison Jaccard similarity, Cosine similarity, and Combined Both of the Data Clustering With Shared Nearest Neighbor Method,"

[13] 2018, volume 5, issue 1, pages 11-18, Comput. Eng. Appl. J.

[14] "Searching research papers using clustering and text mining," 23rd International Conference on Electronics, Communications, and Computing (CONIELECOMP), no. March 2013, pp. 78-81, 2013. [13] E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernandez.

[15] As stated by D. S. Zeimpekis and E. Gallopoulos in their paper "On some document clustering algorithms for data mining," published in Inf. Retr. Boston (14).

[16] "A study on social big data analysis using text clustering," by J. H. Ku and Y. S. Jeong, vol. 7, no. 2, 2018, pp. 1–4.

[17] According to [16] "A Review on Text Mining Techniques," by S. Sathya and N. Rajendran, published in Int. J. Comput. Sci. Trends Technol., volume 3, issue 5, pages 274-284, 2013.

[18] According to [17] M. Allahyari et al.'s 2017 article "A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques,"

[19] "Text Mining : Techniques, Applications and Issues," by R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, was published in 2016 (volume 7, issue 11; pages 414–418).

[20] Reference: [19] B. M. Preethi and P. Radha, "A Survey Paper on Text Mining - Techniques, Applications, and Issues," pages 46-51.

[21] The following is an excerpt from "Clustering Techniques for Text Mining : A Review," written by N. Garg and R. K. Gupta and published in 2016 in the journal Information Science.

[22] Based on the work of G. Loshma, "Nataraj Gudapaty, 2 G Loshma, 3 Dr. Nagaratna P Hegde," 2012, p. 1068-1072, vol. 8491.

[23] For example: [22] M. J. S. Priya, "Clustering Technique in Data Mining for Text Documents," vol. 3, no. 1, pp. 2943-2947,2012.