Ontology Based Text Mining Method Using Cluster Approach

Chinmayee C, S Meenakshi Sundaram*, Keerthana N S, Manikya S, Nitya Hegde M

Department of CSE, GSSSIETW, Mysuru, Karnataka, India

DOI: https://doi.org/10.21467/proceedings.1.1

* Corresponding author email: hodcse@gsss.edu.in

Abstract

In the present world, due to tremendous development in technology, a huge amount of information is available everywhere. Therefore, it is difficult for the users to understand the main content of the entire document as it takes a lot of time. Our project uses the extractive text summarization which uses a method to give the version of summary for one or more file or document. Here we give an approach that maps sentences to nodes of a hierarchical ontology. Ontology explains what exists in a particular domain. For the ontology creation, vocabularies and synonyms are collected. It is used as background knowledge and helps to find the related meaning of the terms which occur in the source documents. Text mining is the technique from which high-quality information is derived from text. Clustering is a significant task. The clustering method groups similar or related terms into a single group. In the first stage, data collection takes place. The preprocessing stage includes stemming and stop words removal.TF-IDF process occurs after which clustering takes place. In the ontology creation, first the determination of the main sub topics of the article of interest is done. Further, the project will extend by giving the refined graph and the summarized text.

Keywords - Stemming, Stop words, TF-IDF, Clustering, Summarization.

1 Introduction

We classify sentences to nodes which have a predefined hierarchical ontology. Each ontology has bag-of-words from a web search. We represent sentences by subtrees that permits to apply measures of similarity and find relations between sentences. The authors use the ontologies for expansion of query, for representing sentences through bag-of-words. The bag-of-tags contain words that are equivalent to ontology concepts. In term-based mapping, the concept of generalization options offered by WordNet relations are exploited to find the concepts that are most informative.

Text mining is the technique from which high-quality information is derived from text. Information of high quality is obtained from consideration of pattern. Text mining includes the process in which the input is structured together with important features and output examining and translation. It helps in the discovery of knowledge that is interesting and summary of data in text documents. Finding the accurate knowledge has always been a challenging issue. Hence serious attention is gained by text mining. Since it has the ability to discover automatically the assets of knowledge concealed in the unstructured text. Text mining tries to discover information which is new and unknown. Techniques from natural language processing and data mining are applied.

Clustering is one of the traditional techniques in data mining. Clustering is an approach in which a set of objects are grouped into subsets or cluster based on their characteristics and similarities. In Clustering, the classification is unsupervised. Example, a web search gives thousand pages as the response to a query. It becomes difficult for users to browse significant information. In methods of clustering, groups are formed for the base word. Similar words are stored in the same group. It separates the documents into group where each group represents a specific topic.

2 Literature Survey

- Madhuri M. Varma *et.al* [1] proposed an Ontology-Based Comprehensive D-Matrix Using Graph Comparison Algorithm. This system comprises the developments of D-matrix from the repair verbatim data.
- Ning Zhong *et.al* [2] presented a pattern discovery technique, pattern deploying, pattern evolving & updating discovered patterns for finding relevant and interesting information.
- Bo Chen *et.al* [3] proposed a framework for adapting text mining models that discovers low-rank shared concept space.
- Jung-Yi Jiang *et.al* [4] proposed a fuzzy similarity-based self-constructing algorithm for feature clustering.
- Jian Ma *et.al* [5] proposed a novel ontology-based text-mining approach to cluster research proposals based on their similarities in research areas.
- Shady Shehata *et.al* [6] proposed a mining model which consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis and concept-based similarity measure.
- Xiuzhen Zhang *et.al* [7] proposed an algorithm for mining feedback comments for dimension ratings and weights, combining techniques of natural language processing, opinion mining and topic modelling.
-

3 System Architecture

Figure1 given below gives the details of the System Architecture used in the Project

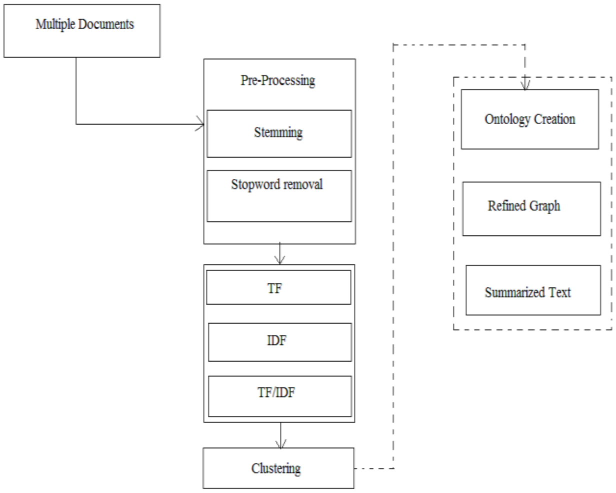

Figure 1: System Architecture

4 Proposed Method

In this project we use extraction summarization approaches to perform the automatic summarization. Extractive methods work by selecting words, phrases or sentences in the original text to form the summary.

- Pre-processing: In the first step splitting of the sentences into words takes place following white space as the separator. The next step is stemming. Stemming is the methodology to get the root of the particular word in the document. The process is continued by removing stop word. Stop word removal is done by comparing each word in the sentence. Examples are articles, conjunctions and prepositions.
- TF-IDF : TF-IDF stands for Term frequency-Inverse document frequency and the tf-idf weight is a weight used in information retrieval and the text mining. It tells how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document.
- TF: Term Frequency tells how frequently a term occurs in a document. Term frequency is often divided by the document length (that is total number of terms in the document).

  TF(t)= (number of times term t appears in a document)/(total number of terms in the document).

  IDF: Inverse Document Frequency, which measures how important a term is while computing TF, all terms are considered equally important. But certain terms, such as "is", "of" and "that" may appear lot of times but they have little importance.

  IDF(t)= log_e(total number of documents/number of documents with term t in it).

tf-idf weight= tf*idf.

- Clustering: Clustering is the process of grouping a similar or related term into a single group. Identification of concepts from text takes place. These are called key concepts of the target domain.

  Steps for clustering are:

  a) A vector of high dimensional concept is given.

  b) Concepts for clustering are generated (that is terms and related terms).

  c) Based on the concepts, the initial clusters are constructed (that is terms and related terms).

  d)The cluster is disjointed to identify the best initial cluster and by the goodness score calculation, the document is kept only in that cluster.

  e) The cluster is built.
- Ontology Creation: For this vocabularies and synonyms are collected. Next, those words are put by the data model of ontology. In the first step, determination of the main subtopics of the article of interest is done. This is obtained by the comparison of words of articles with terms in the ontology. The non-existing words in the ontology are ignored. The number of times the word appears in the ontology is recorded. In the tree structure, each node includes the node's children. If the count of the node increases, the ancestor's count will also be increased. From this type of design, the root of the ontology always gets the highest-level nodes represented by subtopics gets different score. The second-level nodes with higher counts are selected as the main subtopics of the article.

5 Implementation

5.1 Algorithm

Step1: Upload the files.

Step2: File extraction, retrieval of data and create meta data.

Step3: Stemming of the metadata.

Step4: Stop word of the stemmed file.

Step5: Term Frequency(TF).

TF(t)= (number of times term t appears in a document)/(total number of terms in a document).

Step6: Inverse Document Frequency(IDF).

IDF(t)= log_e(total number of documents)/(number of document with term t in it).

Step7: TF/IDF

TF/IDF weight= TF*IDF.

Step8: Clustering groups similar or related terms into a single group.

6 Conclusion

In our project, the initial document corpus is refined into the form of summarization. In this process, we select only the effective features from the refined graph solution. Obtained results are the overall summary of the input document corpus. Summaries provide the readers with condensed versions of the information which is most relevant in the document. It helps readers to assess the document's value without reading the whole document. It acts as content repositories for extracting valuable facts or information.

Acknowledgment

We gratefully acknowledge the help and co-operation offered by Dr. S Meenakshi Sundaram, Project Guide and Professor and Head, Department of Computer Science and Engineering, GSSSIETW, Mysuru. Also, we thank our Project Co-ordinator Rummana Firdaus, Asst. Professor, Department of Computer Science and Engineering, GSSSIETW, Mysuru for providing consistent help and support to carry out this project work. This project has been successfully implemented in the Research Center of GSSSIETW, Mysuru.

References

[1] Madhuri M. Varma et.al They have published "An Ontology-Based Text mining method to construct D-Matrix for fault detection and diagnosis using Graph comparison algorithm" at International Journal of Innovative Research in Information Security in May 2015.

[2] Ning Zhong et.al They have published "Effective Pattern Discovery for Text Mining" at IEEE Transactions on Knowledge and Data Engineering in Jan. 2012.

[3] Bo Chen et.al They have published "Discovering Low-Rank Shared Concept Space for Adapting Text Mining Models" at IEEE Transactions on Pattern Analysis and Machine Intelligence in June 2013.

[4] Jung Yi Jiang et.al They have published " A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification" at IEEE Transactions on Knowledge and Data Engineering in March 2011.

[5] Jian Ma et.al They have published " An Ontology-Based Text Mining Method to Cluster Proposals for Research Project Selection" at IEEE Transactions on Systems, Man and Cybernetics in May 2012.

[6] Shady Shehata et.al They have published "An Efficient Concept-Based Mining Model for Enhancing Text Clustering" at IEEE Transactions on knowledge and Data Engineering in October 2010.

[7] Xiuzhen Zhang et.al They have published "Computing Multi-Dimensional Trust by Mining E-Commerce Feedback Comments" at IEEE Transactions on Knowledge and Data Engineering in Jan. 2007.