Defense Scientific Information & Documentation Centre,

Defense Research & Development Organization

Delhi, India

Ontology-based Document Clustering

January 21, 2019 - February 28, 2019

Introduction

With the wide utilization of the web, an expansive number of printed reports are available over the web. Content information is available wherever on the Web, as big business data frameworks, computerized archives, and individual records. As the extent of content information is expanding at an astounding pace, the taking care of an examination of content information turns out to be essential. Content mining is being produced as an innovation to deal with the expanding volumes of the content information. Diverse content mining functionalities are content grouping, content order, content arrangement.

The utilization of correspondence innovations and data content has expanded broadly. It gives access to a lot of information. Additionally, databases are expanding in volumes step by step and they are of various sorts. In this information, a basic leadership data is shrouded which it should have been broken down so as to acquire learning from the information. Information mining procedures to create designs and get connections from natural information utilizing different investigation instruments. Information mining alludes to separating learning from huge pieces of information. Content Mining is the term used to portray either a solitary procedure or a gathering of procedures in which we get the beforehand unidentified data via naturally extricating data from various advanced information sources. Content databases are additionally expanding because of quickly developing on the web data as electronic archives. The content databases contain data in an unstructured way. Utilizations of Text Mining are progressively vital with the development in the generation of individual and open data with the upgrade of web and internet based life. A lot of data is

available via web-based networking media which is broke down with the assistance of content mining to create significant examples and most recent patterns. Information Mining and its strategies are commonly used to oversee nonnumerical information. Bunching is an information mining strategy that is regularly used to make groups from an extensive number of unstructured information sources which is the nonnumerical information. The bunching strategy has been utilized in huge numbers of the information mining issues, for example, to construct relations from a complex dataset, to discover the relationship between the articles and to make speculations. Bunching applications have been connected to an extensive assortment of regions running from designing, life, and therapeutic sciences, sociology, software engineering, financial aspects, etc.

Content Clustering, an information mining method can be utilized alongside philosophy to gather the comparative computerized information which improves the bunching procedure. A lot of online news information is available on the web. Content Mining can be utilized to extricate comparable news articles from the web. Web-based life is these days a standout amongst the most refreshed types of composing news articles. The expanding volume and accessibility of a lot of online information in web-based life situations give new chances to scientists to screen social and financial conduct. Information Mining can be utilized to separate the comparable news articles from the web and group them based on idea weight and likeness measures.

Related Works

Jian Ma, Wei Xu, proposes an Ontology based Text Mining (OTMM) method to cluster research proposals in a research funding agency. An ontology in the domain of research is created to categorize different research areas. The proposals are then classified into different disciplines and a text clustering algorithm, Self-organized Mapping, is applied to cluster the research proposals on the basis of similarity. After the grouping, the proposals are assigned to the reviewers. Hence this approach reduces the time of grouping the research proposals and assigning to the intended reviewers and promotes efficiency in proposal grouping process.

S. C. Punitha, M. Punithavalli, studied two approaches for text clustering and compared them. First method is based on pattern recognition with semantic driven methods for clustering text documents. Second method is an ontology based text clustering approach. Both algorithms are analysed in terms of efficiency and speed of clustering. Experiments proved that both techniques were efficient in clustering process, but the

performance of ontology based approach was better in terms clustering quality, but a relatively slow speed because of more computations.

E. Alan Calvillo, Alejandro Padilla, proposes a method to cluster research papers by using text clustering. The K-Means algorithm is used to implement the semi-supervised learning clusters to identify and approximate search using as per defined pattern. The limitation of this paper is that it applies semi automatic learning from a knowledge base. An automatic learning process can be applied that can enhance the search from the manipulated texts. Such techniques can be applied to the database knowledge with the help of filter, wrapper and even ontology.

Qiujun LAN proposes an approach to for extraction of news content using similarity measure based on edit distance to separate the news content from noisy information. This paper describes about the accurate extraction of news content from web pages. A backward and forward similarity measure is used based on edit distance method. The algorithms used with this method are less complex with high accuracy and efficiency rate. It is appropriate method to extract news content from noisy data in news web mining.

The Architecture

The proposed system is designed to cluster the text documents. It consists of the following components like preprocessing and ontology. Fig. 1 shows the architecture of the proposed system.

Preprocessing

It involves all processes, methods that are required to prepare data for text mining. It converts data from the original form to machine-readable format before applying feature extraction methods to generate a new collection of documents represented by the concepts. Techniques like stop word removal, stemming and tokenization are involved in preprocessing.

Stop Word Removal

Very often a common word, which would appear to be less significant in selecting a document that would match a user's need, is completely expelled from the vocabulary. Such words are called stop words and the technique is called stop words removal technique. This technique increases effectiveness and efficiency. Example of the stopwords area, is, then, when, etc.

Tokenization

Tokenization is the process of breaking up given character sequences into meaningful words, symbols, or crunches of data while maintaining its security and integrity which can be further used for processing.

Clustering

Text Clustering is the application of the data mining functionality, of cluster analysis, to the text documents. Document or text clustering is an important technique to organize documents. Text Clustering helps to cluster similar kinds of digital documents. This method is used on the web to cluster digital data to enhance the search and to retrieve meaningful lists of the data. Various clustering algorithms are present to cluster similar objects into one cluster and dissimilar objects into a different cluster.

Ontology

Ontology can be considered as a repository of knowledge in which concepts and terms are defined and also the relationships between these terms and concepts are given. It is a set of concepts and relationships that describe a domain of interests and represents an overview of the domain. Ontology makes the knowledge that is implicit for humans, explicit for computers. Hence ontology automates information processing and can improve text mining.

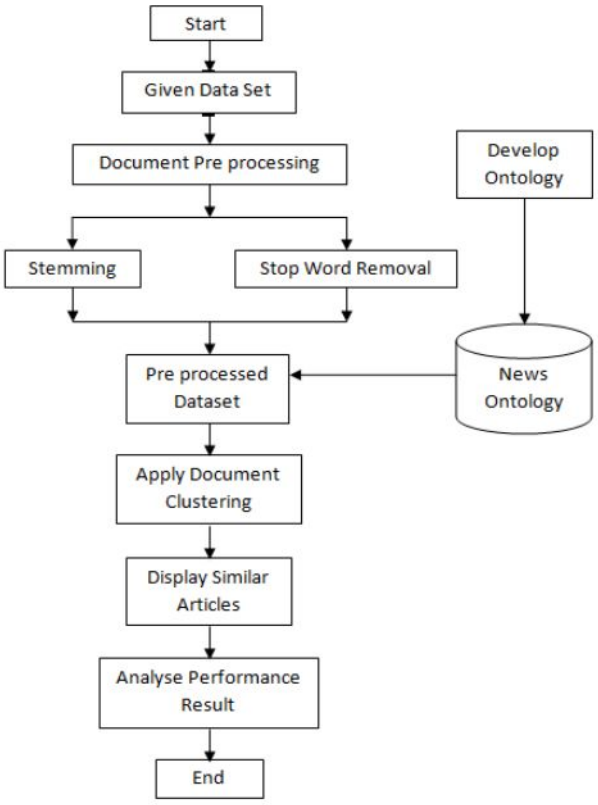

Fig. 1 Architecture of the proposed system

Results

A large collection of text documents are considered as unstructured data. It is very difficult to group the text documents. A dataset is used for the clustering of documents. For this purpose 20 News Group Dataset is used. The dataset consists of a collection of a large number of documents. It consists of a collection of 20000 documents partitioned into 20 different categories. Data from this document collection is taken as input. The documents from this collection are chosen at random for the experiments. Pre-processing techniques are applied to the dataset in order to obtain the pre-processed dataset. Tokenization and stop word removal techniques are applied. Clustering is then applied to the collection of text documents. Fig. 2 shows the clusters of the documents from different categories.

Fig. 2 Cluster Formation

Discussion

Various Text Clustering Algorithms can be used to categorize news articles. The document collection is obtained and pre-processing techniques are applied to the document collection in order to remove stop words and to do tokenization. This would remove unnecessary words from the document collection and would provide a pre-processed dataset. A document clustering algorithm is used for clustering and categorizing the news articles on the basis of topic. Mining similar news articles from the web and applying the ontology-based text clustering algorithm provide clusters of similar news articles. The results of clustering are improved by using the ontology-based clustering algorithms rather than simple clustering algorithms. As future work, such a system can be used to work for big data.

References

Ma, J., Xu, W., Sun, Y. H., Turban, E., Wang, S., & Liu, O. “An ontology-based text-mining method to cluster proposals for research project selection”. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42(3), 2012, ISSN 1083-4427, pp.784-790.

Punitha, S. C., and M. Punithavalli. "Performance Evaluation of Semantic-Based and Ontology-Based Text Document Clustering Techniques." International Conference on Communication Technology and System Design, Procedia Engineering 30, Science Direct, Elsevier 2012, DOI 10.1016/j.proeng.2012.01.839, pp. 100-106.

Agnihotri, D., Verma, K., & Tripathi, P. “Pattern and Cluster Mining on Text Data”. Fourth International Conference on Communication Systems and Network Technologies (CSNT), 2014, IEEE, ISBN 978-1-4799-3069-2, pp. 428-432.

Calvillo, E. A., Padilla, A., Munoz, J., Ponce, J., & Fernandez, J. T. “Searching research papers using clustering and text mining”. In International Conference on Electronics, Communications and Computing (CONIELECOMP), 2013, IEEE, ISBN 978-1-4673-6156-9, pp. 78-81.

Qiujun, L. 2010. "Extraction of News Content for Text Mining Based on Edit Distance", Journal of Computational Information Systems, 2010, pp.3761-3777.