EAI Endorsed Transactions
on Energy Web and Information Technologies
Research Article EAI.EU

Conceptual Semantic Model for Web Document Clustering Using Term Frequency

Dr.N.Krishnaraj[1,*],Dr P Kiran Kumar[2] and Sri K Subhash Bhagavan[3]

[1,2]Professor, Sasi Institute of Technology and Engineering
[3]Assistant Professor, Sasi Institute of Technology and Engineering

Abstract

Term analysis is the key objective of most of the methods under text mining, here term analysis either refers to a word or a phrase. Determination of the documents subject is another important task to be performed by the semantic based method; this is done by identifying those expressions that resemble the semantics of a sentence. This model in general is called as the mining model and it is exclusively used to identify either the words or the expressions in a document on each and every specific sentence, this identification can also be done at the core level. As far as a group of documents is concerned the proposed method is capable of identifying the similar concepts among them; this identification is done by analysing the sentence semantics among the documents. The prime focus is to improve the quality of the web document clustering method, this is done by analysing the semantics of the sentences efficiently and thereafter organising the same effectively.

Keywords: Clustering, Semantic Model , Text Mining, Term Frequency

Received on 06 July 2018, accepted on 16 August 2018, published on 12 September 2018

3rd International Conference on Green, Intelligent Computing and Communication Systems - ICGICCS 2018, 18.5 - 19.5.2018, Hindusthan College of Engineering and Technology, India

doi: 10.4108/eai.12-9-2018.155744

*Corresponding author. Email:krishnaraj@sasi.ac.in,kiran@sasi.ac.in,suba@sasi.ac.in

1. Introduction

1.1 Semantic Web

In general, semantics is the study of meaning. If a computer understands the semantics of a document, it doesn't just interpret the series of characters that make up that document, but help to separate meanings from data, document content, or application code, using technologies based on open standards. The current WWW has a huge amount of data that is often unstructured and usually only human understandable. Semantic Web is an extension of current Web which offers to add structure to the present Web. The Semantic Web [3] aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way. The goal of the Semantic Web is to develop enabling standards and technologies designed to help machines understand more information on the Web so that they can support richer discovery, data integration, navigation, and automation of tasks. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

1.2 Semantic Web Architecture

The development of the Semantic Web proceeds in layers, one above another allowing for a more standardized way of developing. As it is being built on existing technology it allows developers to roll out parts of technology and implementing them without realizing the full capabilities of the Semantic Web. The functionality of each layer with reference to the above layered architecture is represented below with Semantic Web Layered Architectureis shown in Figure 1 [4].

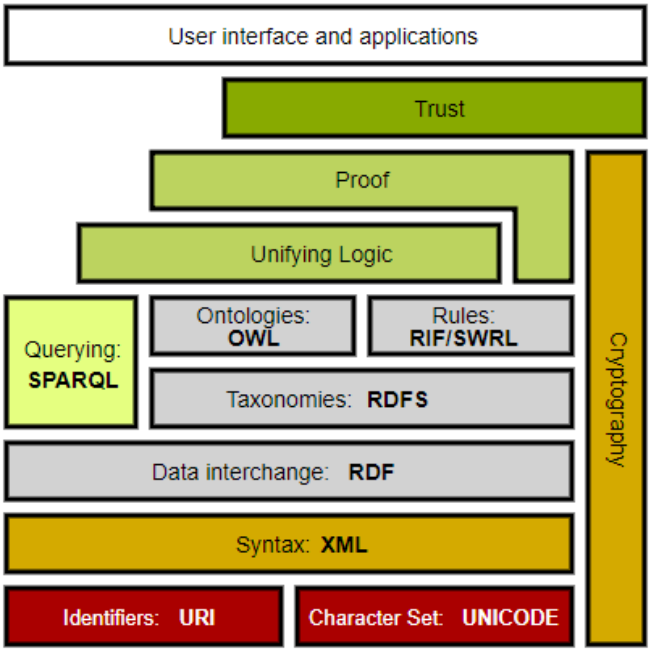

Fig.1 Semantic Web Layered Architecture

The entire world today is in search of ways and means where they can access information about anything and everything just by sitting at their own comfortable locations, the internet thus makes this accessing search convenient for users at any nook and corner of the world. It acts as a fundamental resource provider and thus seen as a major data storage component. Vast storage of information makes it an overloaded component. This feature makes it difficult for a normal user to search for and arrive at precise information's; also this difficulty is due to its constantly changing characteristics. To overcome this problem many applications with built in features that provides the required flexibility for a normal user has been introduced in order to make the search operations user friendly.

The interaction between a server and a user normally resembles a dialogue that happens between two individuals, the process begins by transmitting a request message to the server, normally referred to as the query, on reception of the query message the server would transmit back a suitable reply to the user that is normally referred to as the response. The search engine actively functions in this process by means of verifying the query message word by word carefully and then filters out the suitable response message that matches the query. This response usually takes the form of a web page. Search engines usually correspond to URLs that are quite difficult to search. There arises another problem here, where the users with insufficient domain knowledge may not be aware of the exact key words required for the search, this may lead to the openings of unwanted or totally unrelated web pages that might totally confuse the user and make his/her consolidation process difficult. These problems or difficulties has motivated the idea of introducing suitable methods that would assist the user in their tracking task, in a way of enabling them to precisely track and organise their web documents as per their requirements.

The ultimate aim of this method is to trace responses that exactly match the requirements of the user. This goal can be achieved by incorporating the Document clustering method that plays an efficient role in precise tracking. Additional improvisations have been made on this document clustering method in the later stages for improving its significance further; various applications have also been introduced in order to make information processing and its management much more efficient.

Most text clustering techniques are based on words and/or phrases weights in the text. Such representation is often unsatisfactory because it ignores the relationships between terms, and considers them as independent features.

2. PREVIOUS WORK

The documents involved in the text clustering method are usually split into clusters or groups. The documents present in each group may represent some common properties with the other documents in the same group. Each and every group would hold a topic of their own and it is based on this topic that the groups are differentiated from each other. The documents clustering techniques are based on the concepts of the "Vector Space Model (VSM)". This model is used as a data representation model in the text classification and clustering fields. As far as the "Vector Space Model" is concerned the documents are represented as a set of feature vectors. Terms are considered as the most important component in a document, this is well stressed by the feature vectors, these vectors hold a term called as the term weight (i.e. term frequency) that essentially represents their importance in a document. The homogeneity among the documents is evaluated with the help of similarity measures (e.g. Cosine measure, Jaccard measure).

The text clustering process is a collection of various methods like conceptual clustering, decision trees, neural nets, rule-based system, statistical analysis, clustering based on data summarization, and inductive logic programming.

Representation of a document is usually done by means of selecting suitable features, features are best known to represent it, in a document clustering method this task of selecting the features is considered important and is given more priority. This selection process affects the clustering result in a tremendous way. Accuracy of the results produced by the clustering process depends on the value of the weight terms; these terms are evaluated for the feature factors. The weight terms must be evaluated accurately as these terms have a great impact on the final clustering results.

3. PROPOSED WORK

The work done in the previous methods for the purpose of clustering the documents have been discussed so far and these strategies are incorporated for clustering the documents that are available on the system. In the proposed work web document usage will be given more importance rather than the plain text documents. Figure 2 represents the dependence on the concept based mining model, this model is meant for an analysis purpose, where the analysis is done at the various levels such as the sentence, document and core levels in order to identify the important terms. Another advantage of this model is that it helps in improving the quality of the cluster and this improvement is done with the help of the "concept-based similarity measure".

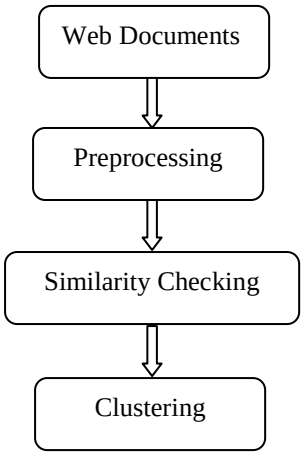

Fig. 2: Flow Diagram of Semantic-Based Model for Document Clustering

Web documents are fed as inputs to the proposed model; these inputs are processed suitably by the model, which then produces the cluster outputs which are of good quality. This good quality clusters thereby improve the quality of the search engines. Well defined sentences make up a good document. Every sentence in a document would be labelled; this is done by the semantic roll labeller that automatically performs the labelling task. The sentences are labelled as per the "Prop Bank notations" and the outputs are labelled as per the "verb argument structure". The "concept-based model" is used for the purpose of analysing the labelling process's results which is done at the sentence, document, as well as at the core levels. This model finally identifies the concepts based on the labelled terms.

3.1 Pre-processing Stage

Stage 1: Sentence Separation

Every sentence in a document is made up of terms, identifying or analysing the terms in a document involves a simple procedure of determining the relationship among the verbs and their arguments. Once this term identification process is over the next step would be to understand the need of every term in a sentence as these terms involve themselves in building the sentence semantics.

As already seen, the semantic roll labeller that is meant for assigning labels to each sentence in the proposed work is utilised for determining the relationship of the terms towards the sentence semantics and best determines the one that has a very close meaning of the sentence under consideration. The verb argument structure is used to characterize the sentence-semantic structure and it makes a connection among the arguments of the input query with their corresponding semantic-role.

Determination of the arguments of verbs is another important task in clustering methods; this is done by the "Support Vector Machines (SVMs)". Arguments are usually classified based on their semantic roles. SVMs perform this classification. The roles are as follows, Agent, Theme, and Goal. Comparing the performance of the earlier classifiers with that of the SVMs, it is seen that the SVMs have an improved performance result.

Stage 2: Stop words and stem words identification and removal

Removal of stop words would fall under the data cleaning process, this cleaning process is done to exclusive remove the stop words. This cleaning process is activated immediately after the completion of the verb argument structure formation. Porter Stemmer algorithm is used for stemming the words. The terms that are remaining after the elimination of the stop words and the stem words would be essentially called as the concept.

3.2 Semantics-Based Analysis

Both at the document and at the core level analysis of the concepts on each and every sentence is done in order to get a clear idea about the same. Analysing the entire document in terms of a single term fashion would not achieve the goal of understanding the entire concept; hence the "semantics-based analysis" method is incorporated in order to make a detailed investigation about the concepts on each and every sentence.

Concept Analysis at Sentence Level:

Each and every labelled term would carry a frequency parameter (or concept), this parameter is identified by using a frequency measure that is also known as "conceptual term frequency (ctf)" which is meant for analysing the concepts at sentence level.

Concept Analysis at Document Level:

The "concept based term frequency (tf)" is used for analysing the concepts at the document level and the same is computed by means of calculating the number of occurrences of a labelled term in a verb argument structure of a sentence. At this point, the 'tf' of a concept is also analysed at the document level.

Concept Analysis at Corpus Level:

Discriminationamong the documents is done so as to have a better understanding of the concepts. This discrimination process is executed by the concepts or labelled terms and the one that makes the best is filtered out by using the "concept-based document frequency (df)". This 'df' value clearly portrays the discrimination task. The 'df' factor is considered as a global measure as it clearly analyses the concepts at the corpus level.

3.3 Concept-Based Document Similarity

The measure that is based on the concept based similarity is influenced by the following important features: The first step is the consideration of the concepts, identification of the semantics of a sentence by the labelled terms that is eventually produced by the pre-processing steps are considered here as the concepts. Second, the number of occurrences of a labelled term in a verb argument structure of a sentence helps to identify the importance of the term towards the sentence-semantics, in addition to the document's main subject.

Third, while computing the similarity between the documents the concept-based document frequency 'df' is used for the purpose of differentiating the documents from one another.

3.4 Clustering Method

Suffix tree clustering (STC) algorithm is incorporated in the proposed method. In this method the documents are treated as a set of sequential words. STC performs the task of identifying common phrases in a group of documents, this identification or execution time is usually linear.

STC comprises of three logical steps: a. document cleaning, b. identification of base clusters, and c. combining these base clusters to form a common root cluster. STC is seen as a fast document clustering algorithm, so it will have a significant role in clustering web documents.

4. CONCLUSION

The proposed work brings in an association among the text mining and the natural language processing disciplines. The "semantic-based mining model" that is incorporated in the proposed work considerably showcases a desirable improvement in the quality of the clusters. An improvement in the clustering result is due to the proper identification of the sentence semantics structure in a document, this determination task is evident and does plays an important role in the quality of the results. The quality of the output clusters obtained by this method may possess considerable improvements when compared with that of the traditional single term-based approach.

REFERENCES

[1] The ACM Conference on Information and Knowledge Management:Shah, U., Finin, T., Joshi, A., Mayfield, J., & Cost, R. (2002), "Information retrieval on the semantic web", , November 24.

[2] Springer-Verlag Berlin Heidelberg:McCuaig, J. (2011), "The Semantic Web. In: Essential Software Architecture", DOI 10.1007/978-3-642-19176-3_12.

[3] Semantic Grid – The Convergence of Technologies: Stumme.G, Hotho.A, Berendt.B(2006) "Semantic Web Mining: State of the art and future directions", Web Semantics: Science, Services and Agents on the World Wide Web 4(2) 2006 124-143.

[4] Journal of Universal Computer Science:Eddie Moench, Mike Ullrich, Hans Peter Schnurr, Juergen Angele(2003), "Semantic Miner: Ontology Based Knowledge Retrieval", Vol.9, No.7 682-696.

[5] Oikonomakou, Nora, and MichalisVazirgiannis, (2005). "A Review of Web Document Clustering Approaches." Data Mining and Knowledge Discovery Handbook.

[6] IEEE Trans.Pattern Analysis and Machine Intelligence:P. Mitra, C. Murthy, and S.K. Pal, (2002). "Unsupervised Feature Selection Using Feature Similarity" 24(3): 301-312.

[7] D. Gildea and D. Jurafsky(2002). "Automatic Labeling of Semantic Roles," Computational Linguistics, 28(3): 245-288.

[8] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky(2004). "Shallow Semantic Parsing Using Support Vector Machines," Proc. Human Language Technology/North Am. Assoc. for Computational Linguistics (HLT/NAACL).

[9] IEEE Trans.Pattern Analysis and Machine Intelligence:H. Jin, M.-L. Wong, and K.S .Leung (2005). "Scalable Model-Based Clustering for Large Databases Based on Data Summarization" 27(11): 1710-1719.

[10] Machine Learning:S. Pradhan, K. Hacioglu, V. Krugler, W. Ward,J.H. Martin, and D. Jurafsky (2005). "Support Vector Learning for Semantic Argument Classification" 60(1-3): 11-39.

[11] IEEE Trans. Pattern Analysis and Machine Intelligence: R. Nock and F. Nielsen(2006). "On Weighting Clustering," 28(8): 1223-1235.

[12] Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM): S. Shehata, F. Karray, and M. Kamel(2006). "Enhancing Text Clustering Using Concept-Based Mining Model "

[13] Issues in Informing Science and Information Technology:Samuel Sambasivam, and Nick Theodosopoulos, (2006). "Advanced Data Clustering Methods of Mining Web Documents", 3: 563-579.

[14] Special Issue in Journal of Advanced Research in Dynamical and Control Systems Thilagavathy R, Sabitha R, Vol 12, 526-545,2017.

[15] Asian Journal of Scientific Research, A. Kousar Nikhath and K. Subrahmanyam, 2018. Conceptual Relevance Based Document Clustering Using Concept Utility Scale , *11: 22-31.*