Semantic Analysis Based Approach for Relevant Text Extraction Using Ontology

Poonam Chahal, Manav Rachna International University, Faridabad, India
Manjeet Singh, YMCA University of Science and Technology, Faridabad, India
Suresh Kumar, Manav Rachna International University, Faridabad, India

ABSTRACT

Semantic analysis computation is done by extracting the interrelated concepts used by an author in the text/content of document. The concepts and linking i.e. relationships that are available among the concepts are most relevant as they provide the maximum information related to the event or activity as described by an author in the document. The retrieved relevant information from the text helps in the construction of the summary of a large text present in the document. This summary can further be represented in form of ontology and utilized in various application areas of information retrieval process like crawling, indexing, ranking, etc. The constructed ontologies can be compared with each other for calculation of similarity index based on semantic analysis between any texts. This paper gives a novel technique for retrieving the relevant semantic information represented in the form of ontology for true semantic analysis of given text.

KEYWORDS

Concepts, Ontology, Parsing, Semantic, Similarity, Syntactic Analysis

INTRODUCTION

Natural Language Processing is the field which works in the syntactic and semantic analysis of the text. The major issue concerned related to the field of Natural Language Processing (NLP) is keyword-sense disambiguation related to text (Seo et al., 2004). The disambiguation and visualisation of words is difficult as the identification of the relevant word or semantic understanding of the same word is a critical task (Wagh & Kolhe, 2011). The semantic understanding of any type of information which has multiple meaning have many issues as the volume of the information is constantly increasing every day. For example, the word “Bank” can have numerous meanings such as financial body, watercourse bank, heap into a mound, blood bank etc. To resolve the issues of semantic analysis of text many methods have been designed and developed but the applications of these methods in the area of information retrieval are not up to the level of user expectations. This paper deals with the issue of extraction of relevant concept from a given text so that semantic information associated with the text is retrieved. Our method of analysis the text/content of a document extracts the relevant interrelated concepts so that the summary of the same text/content can be created. Finally, obtained relevant summary is represented in the form of ontology and thus further used for computation of semantic similarity between given texts. Thus, the proposed approach given in this paper make the use

DOI: 10.4018/IJIRR.2017100102

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
of knowledge structure i.e. ontology which is a prescribed arrangement of common conceptualization (Gruber, 1995).

In this paper, the work done by the various researchers in the area of relevant text retrieval from a document depending upon semantic analysis is discussed. Next, the proposed approach is given which aims to extract the relevant text having the most semantic significant information from the large text. The proposed technique is also evaluated with some of the already presented approaches of semantic analysis of document and the results showing the superiority of the proposed approach is also shown. Finally, we talk about the conclusion and potential scope of the given text extraction approach.

RELATED WORK

(Pal & Saha, 2014) have given an algorithm named simplified Lesk which is based on an unsupervised approach. This algorithm helps in extracting the summary from a large content of a document. The construction of summary is done for obtaining the important sentences with the help of conventional online dictionary. These significant sentences are assigned weights according to the importance of the content written by an author. Further these assigned weights are evaluated using the steps of the proposed simplified Lesk algorithm. This simplified Lesk algorithm only considered the keywords by using the WordNet which is not enough to extract the meaningful information from a document. So, the simplified Lesk algorithm is modified which considers the semantics extracted from conventional online dictionary like WordNet.

(Kulkarni & Apte, 2014) has given lexical chain computing algorithm in which the division of original documents is done into sentences using segmentation. Then a map is constructed for each sentence using four types of relations i.e. synonym, hypernym, hyponym, meronym. The distance of each word with another related word is calculated and a lexical chain is constructed which is assigned weight accordingly. The chain having highest weight is selected as the longest chain and the relevant text extraction is done depending on the relevant lexical chains obtained.

(Patil et al., 2004) used the statistical approach for extracting the relevant information from a document. The given approach can be used in application for retrieving the semantic association from a text depending on the query from the user in the form of keyword. In this paper, the process of extracting relevant information in the form of summary is done in various phases like extraction of data items, generation of query, documents gathering, creation of summary, and arrangement of summary to any user of a query.

(Hennig et al., 2008) has discussed an approach in which keywords extracted from the content of a document are mapped to the nodes of the knowledge structure i.e. ontology. This mapping is done to give the semantic representation to the content of document which further helps in extracting the semantically relevant information from a document. The classifier that maps sentences to the classification is very flexible and it is also not linked to a specific domain as it is trained by means of search engines. The SVM classifiers are skilled so that they can recognize summary sentences by using the ontology-based features.

(Gaona et al., 2009) has given a means for semantic gain computation depending upon sense of each word from the content of the information. However, the computation of regularities of co-occurrences among gloss and its context requires various amount of information stored. The computation of regularity is through a query specified to the search engine and further choosing intellect which make use of the weight.

(Thiagarajan et al., 2008) gives the similarity computation based on semantic analysis using ontologies. The authors discussed that generally, a web document is considered as set of words which is also called as Bag of Words (BOW). In contrast to the BOW approach the authors considered the document as set of concepts called as Bag of Concepts (BOC) as the keyword approach is lacking of intelligence. So, to extract the semantics from a document the process of spreading is used so
Related Content

Virtual Community of Practice Ontocop: Towards a New Model of Information Science Ontology (ISO)
www.igi-global.com/article/virtual-community-practice-ontocop/58891?camid=4v1a

The Use of Text Mining Techniques in Electronic Discovery for Legal Matters
Michael W. Berry, Reed Esau and Bruce Kiefer (2012). *Next Generation Search Engines: Advanced Models for Information Retrieval* (pp. 174-190).
www.igi-global.com/chapter/use-text-mining-techniques-electronic/64425?camid=4v1a

Chronological Ordering Based on Context Overlap Detection
Mohamed H. Haggag and Bassma M. Othman (2012). *International Journal of Information Retrieval Research* (pp. 31-44).
www.igi-global.com/article/chronological-ordering-based-on-context-overlap-detection/90440?camid=4v1a
Dependency Graph Based Detection of Semantically Equivalent Questions in Online Forums
www.igi-global.com/article/dependency-graph-based-detection-of-semantically-equivalent-questions-in-online-forums/217483?camid=4v1a