IT.ENG.4

CLUSTER BASED NATURAL LANGUAGE QUESTION ANSWERING FOR E-RESOURCES IN A LIBRARY

M. A. C. Akmal Jahan[1], M. A. C. Jiffriya[1], Roshan G. Ragel[2], Sampath Deegalla[2]

[1]Post Graduate Institute of Science, University of Peradeniya
[2]Department of Computer Engineering, Faculty of Engineering, University of Peradeniya

Processing questions and question answering (QA) based on a pool of text documents is an information retrieval mechanism which responds to queries in natural language by returning brief answers. QA involves (1) understanding user questions which are asked in a natural language through a Natural Language Processing technique and (2) categorizing the documents based on data mining for fast and real time behaviour of the system. The work presented in the abstract targets the second issue for an e-library system of a university.

In a library, there are a large number of resources in electronic format similar to printed material such as e-books, e-journals, etc. Most of the users do not know the mechanism of selecting the appropriate e-resource for their reference. When a user wants a brief answer for a question or most appropriate source to extract the answer, he should search the entire source documents in a particular category. Such an approach will incur a significant amount of time mostly failing the real time requirements and sometimes the required answer may not be available in the documents.

The problem of unacceptable time delay may be overcome by introducing an e-library management system in a university and accessing information by processing natural language questions through the system, which stores all the electronic resources. This is achieved by narrowing down the search space by looking at the index or abstracts of the resources and looking for answers by matching required question with narrowed down resources.

For this purpose we have applied an approach of cluster based document retrieval in which matching a query is performed against clusters of documents instead of individual documents. Therefore, we need to find an appropriate clustering algorithm for document clustering which is accurate and efficient for the problem domain. We have tested data from the given domain with K-Means, Hierarchical, EM and Cobweb clustering algorithms available in a well-known data mining tool named WEKA. Numbers of incorrectly clustered instances of the documents and the computational time for clustering have been measured for each algorithm. According to the results, K-Means algorithm showed 74% of accuracy with six clusters and performed faster than the other clustering algorithms. Since the K-Means algorithm shows higher performance on document clustering it was selected for constructing clusters for the e-resource pool.