2012 IEEE Fourth International Conference on Technology for Education

IPedagogy: Question answering system based on web information clustering

Rivindu Perera
Department of Computer Science
Informatics Institute of Technology
Colombo 06, Sri Lanka
rivindu.perera@hotmail.com

Abstract— **As with the excessive information growth in the web, retrieving the exact segment of information even for a simple query, has transformed to a difficult and resource expensive state. Specially, in e-learning domain it is vital to search knowledge frequently and focusing on a limited well defined search space. IPedagogy is a question answering system which works with natural language powered queries and retrieve answers from selected information clusters by reducing the search space of information retrieval. In addition, IPedagogy is empowered by several natural language processing techniques which direct the system to extract the exact answer for a given query. System is evaluated with the use of mean reciprocal rank and it is noted that system has 0.73 of average accuracy level for 10 sets of questions where each set is consisted of 35 questions.**

Keywords-question answering system; information clustering; information retrieval; natural language processing, mean reciprocal rank; e-learning

I. INTRODUCTION

Research suggests the gradual shift towards the usage of information clustering techniques with the inherent features of question answering systems to support e-learning. This approach, models the goal of the research as a hybrid approach that can be presented with information clustering. Natural Language Processing (NLP) techniques such as Named Entity Recognition (NER) and Relation Extraction (RE) [1] are engaged in an environment where several synchronous and asynchronous communications take place in order to achieve the objective of outputting the answer. But among all, clustering plays a significant role in reduction of search space and thus decreasing the work load of resource expensive NLP techniques.

There are gaps in usage of the clustering process effectively to reduce the information loss. For an example, usage of a clustering process which does not assign a high priority level to the cluster labels will eventually come up with invalid or inaccurate answers. Another notable point in current question answering systems is that once the process of information extraction is terminated the acquired information is not justified or validated to certify the correctness.

The paper, therefore, seeks to explore and evaluate the usage of information clustering in web based question answering systems through the developed prototype, IPedagogy.

II. BACKGROUND OF THE STUDY

A. Information clustering

The rationale for information clustering in information retrieval systems such as question answering systems is that resulting clusters can be used as potential answer sources. Park's [2] seminal work in cluster based information retrieval investigated the term-frequency inverse document-frequency weighting based clustering. But with the same schema and idea Yan and Li [3] moved in depth of clustering with Spherical K-means (SK-means) clustering algorithm.

Nevertheless, the approach presented by Yan and Liu is more accurate when comparing with the Park's approach and it is identified the advantage of using criterion function [4] in the process though Yan and Liu have not considered.

Han et al. [5] express the usage of Suffix Tree Clustering (STC) with web based information resources by condensing the imperative section of the resource such as keywords which need to be focused. The technique incorporated in this research concentrates term based search and text snippet extraction.

B. Named entity recognition

Entity extraction is not longer considered as an uncomplicated classification problem. Reason behind this is that as emphasized by this research, information growth in the web and the demand.

Maximum Entropy (MaxEnt) approach and Hidden Markov Model (HMM) integration proposed by Biswas et al. [6] is one of the emerging approaches recently appeared in the area due to the simplicity, accuracy and also the effectiveness of the solution presented.

C. Relation extraction

Identifying the subject, verb and object can be considered as a basic linguistic method, but that can be seamlessly integrated to several computational linguistic systems.

Several approaches like Conditional Random Fields (CRF) and MaxEnt with parse tree generation [7] can be considered in the relation extraction context. Among them, MaxEnt with parse tree generation mechanism which is a widely used in the NLP based systems and as a discriminative classification method, significant productivity is also highlighted.

978-0-7695-4759-6/12 $26.00 © 2012 IEEE
DOI 10.1109/T4E.2012.48

IEEE computer society