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.

III. METHOD

A. Research design

This research employed a pretest-posttest quasi experimental design to examine the accuracy of answers through cluster analysis. Therefore, cluster labels with the terms identified in the questions are considered by analyzing the matching distance and also the questions are formed by considering several categories of knowledge.

B. Algorithms and techniques

Summary of algorithms used is shown in Table I below.

TABLE I. SUMMARY OF ALGORITHMS

Process / method	Algorithm(s)
Term extraction	MaxEnt
Information clustering	STC
RE	MaxEnt
NER	MaxEnt + Ontology
Text matching	Levenstein + Smith Waterman

Clustering process where STC is used is designed in such a way that the algorithm can be incorporated with the search result set in a more generic way. To implement this generic way of communication Extensible Markup Language (XML) is used.

Named entity recognition process is powered by MaxEnt and ontology based approach which enables the application to dive deep in to the document. Currently, IPedagogy is powered by 28 different entity types such as person, country, automobile, operating system, holiday etc.

Text matching is performed via combination of Levenstein and Waterman algorithms [8] where combination designed to identify the matching answer which has lowest distance from object type of the relation and highest matching precision using projection of object as shown below in Fig. 1 using an example scenario.

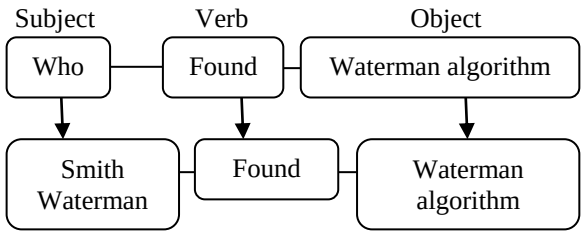

Figure 1. Example of answer projection

IV. RESULTS

With 10 sets of questions where each set was consisted of 35 questions, were used in the testing phase of the IPedagogy and the key concept used here was the Mean Reciprocal Rank (MRR) calculation for each set. As shown below in Fig. 2, IPedagogy gained 0.73 average MRR value while for START [9] question answering system with same question set, it is noted as 0.70. Also lowest individual MRR value of IPedagogy is identified as 0.6 and for START, it is noticed as 0.52.

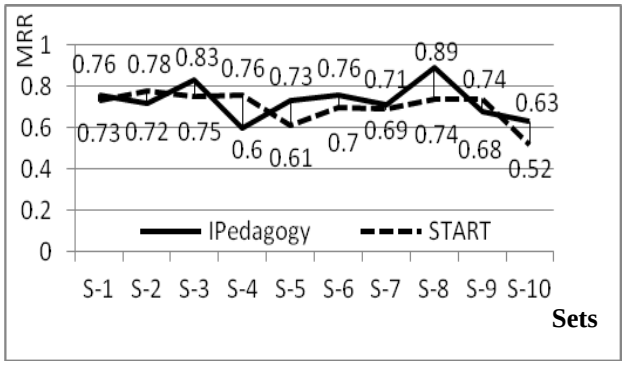

Figure 2. Result of evaluation using MRR

V. DISCUSSION AND INTERPRETATION

Research findings suggest that usage of information clustering has affected the question answering on a more positive way by showing high level of accuracy but still there are points which need to be addressed by better algorithms. For an example, during testing it is identified that text matching algorithm usage has caused an information loss preventing the answer to be extracted by ignoring matching text in the context. Another key identification is that to maximize the cluster analysis space to identify more clusters from the web search result list. Currently, IPedagogy is able to identify 68% of possible clusters from a result list.

When considering the 0.73 MRR in the testing phase it is noted that this application can be successfully used in several e-learning applications which will enable the students to get answers with high accuracy than what they get through a simple web search.

REFERENCES

[1] E. Marsh and D. Perzanowski, "MUC-7 evaluation of ie technology:Overview of results", Message understanding conference, 2005.

[2] J.M. Park, "Intelligent query and browsing information retrieval (IQBIR) agent". IEEE international conference on acoustics, speech and signal processing, Seatle.1998.

[3] H. Yan, C. Lin and B. Li,"A SVM-based text classification method with ssk-means clustering algorithm", International conference on artificial intelligence and computational intelligence, 2010.

[4] S. Na, L. Xumin and G. Yong, "Research on k-means clustering algorithm: an improved k-means clustering algorithm", Third international symposium on intelligent information technology and security informatics. Jinggangshan. 2-4 April 2010. New York, pp.63-66. 2010.

[5] H. Wen, N. Xiao and Q. Chen, "Web snippets clustering based on an improved suffix tree algorithm". Sixth international conference on fuzzy systems and knowledge discovery, Guangzhou, 2009.

[6] S. Biswas, S. Mohanty and S. P. Mishra, "A Hybrid Oriya Named Entity Recognition System: Integrating HMM with MaxEnt". Second international conference on emerging trends in engineering and technology , Nagpur, 2009.

[7] S. Zhang, J. Wen, X. Wang and L. Li, "Automatic entity relation extraction based on maximum entropy". Sixth international conference on intelligent systems design and applications, Jinan , 2006. pp.2-4.

[8] B. Soewito and W. Ning, "Methodology for evaluating string matching algorithms on multiprocessor", International conference on computer systems and applications, Doha, 2008, pp.2-4.

[9] Katz, G. Borchardt and S. Felshin, Natural language annotations for question answering. 19th International Florida artificial intelligence research society conference, Melbourne Beach, 2006, pp.1-4.