

**PERSONALIZED QUESTION ANSWERING SYSTEM BASED
ON IR AND QUERY REFORMULATION**

Thesis

Submitted to

Babasaheb Bhimrao Ambedkar University

(A Central University)

Lucknow

**BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY**



LUCKNOW
प्रज्ञा शील करुणा
ESTABLISHED 1996

For the degree of

Doctor of Philosophy

In

Computer Science

By

VAISHALI SINGH

Under the Supervision of

Prof. SANJAY KUMAR DWIVEDI

Department of Computer Science

Babasaheb Bhimrao Ambedkar University

(A Central University)

Lucknow-226025, India.

Abstract

In recent times, Question Answering (QA) has gained enormous attention from the research community. Especially, during the last two decades, a number of efforts have been put to enhance the performance of QA in open domain and in restricted domain as well. The technology of QA systems lies at the cross-road of NLP (Natural Language Processing), Information Retrieval, Machine Learning, Pattern Learning etc. The task of question answering can be defined as finding a specific text segment in response to a given question. It is enhanced form of information retrieval resulted from understanding complex syntactic and semantic relationships of user's question implementing linguistic techniques, machine learning or pattern matching. It is this perspective with which we initiate our study in question answering by analyzing and presenting different approaches used in the research area.

Meanwhile, the present trend in question answering is focusing more on open domain than restricted one. An open domain QA system however lacks the capability to deal with specific questions of restricted domain as no constraint has been imposed on user's vocabulary. Also, practically it is difficult to develop external resources like thesaurus, dictionary, and knowledge resource etc. for an open domain. Therefore, task of developing QA systems for specific domain has also been accelerated in recent past. Number of restricted domain system have been developed for tourism, biography, geography, weather and also for e-learning but none of the system focuses specifically on the question related to higher education domain.

With this background, we propose a question answering system which is restricted to answer only questions on higher education domain. We have utilized two types of information resources i.e., QA corpus and Web to enhance performance and reliability of the system. In due course of developing the QA system, we have worked on increasing accuracy of question class and focus word detection which enhances the primary understanding of user question. We have also worked on notion of sub-division of local information resource i.e., QA corpus depending on domain based constraints which reduces the overhead of searching whole corpus while answer extraction. Our proposed QA system makes extensive use of patterns to balance a trade-off between complexity and performance instead of using deep linguistic analysis. Pattern matching is used as primary approach for question classification and query reformulation which is later on complemented with Support Vector Machine (SVM) and Structured Query Reformulation (SQR) respectively. We have focused on exploiting maximum information from the previous modules

to the next one. This is the reason, why we have again utilized pattern corpus as an evidence for SQR. Similarity metrics have been used for answer retrieval from both the resources. However, extraction from Web implements some additional metrics to filter precise answer string from large data. The issue of personalization has also been addressed by developing user model based on entities and their attributes observed in user's previous question. Finally, the modules of standard QA system and personalization are integrated to present the unified framework. The experiments results showed satisfactory MRR (Mean Reciprocal Rank) and proved it is feasible for the viewpoint of practicality using the proposed method for restricted domain.