Design of a Novel Search Engine for Prospective Question Answering

Rosy Madaan, Department of Computer Engineering, YMCAUST, Faridabad, India
A.K. Sharma, Department of Computer Science & Engineering, BSAITM, Faridabad, India
Ashutosh Dixit, Department of Computer Engineering, YMCAUST, Faridabad, India

ABSTRACT

Question answering offers a more intuitive approach to information processing. A number of approaches have been used for answering questions. In this paper, we propose a question answering system that uses blogs as its source of information. The system deals with crawling blog pages, summarizing them, indexing and then ranking the summarized content. The user asks a question and gets answer(s) in response. The answer(s) obtained are better as compared to those provided by the existing QA systems that use the general web pages for the purpose of answering. The experimental results show that the proposed system has shown promising results and the responses given by the system are better than those given by the existing QA systems.

Keywords: Blog, Crawler, Indexer, Question Answering, Ranking, Search Engine, Summarization

1. INTRODUCTION

Search engines, in Arasu et al. (2001) and Brin et al. (1988) provide users a way to find information in an organized fashion. A search engine uses special software called a crawler, given in Burner et al. (1997) and Yates et al. (2004) to go through the entire Web and download Web pages using a list of seed URLs. The downloaded Web pages are then stored in a repository and are scanned to extract the URLs existing in them. These URLs are further added to the list of seed URLs for further crawling. After crawler downloads pages, they are then stored in repository and are then passed to another program called indexer, given in Arasu et al. (2001) and Brin et al. (1988) for indexing. Indexer (in nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf) maintains an index that maps keywords to the documents in which they appear. The index is then searched for the answering user’s query.

Yates et al. (1999) discussed that the search engines retrieve the information in the form of documents. In case if the user needs an answer to a question, the user has to go through the entire document to extract the answer to his

DOI: 10.4018/ijirr.2014040102
question, which is a time consuming process. So there is a need to build a system which is capable of taking user’s question as input and returns answer(s) as the result. Hence, the need of Question answering system arises.

Question answering offers a more intuitive approach to information processing. Given a collection of documents and a natural language question posed by user, a question answering system attempts to find the precise answer or at least a portion of the text in which the answer appears. Architecture of a general Question answering system, in Zheng et al. (2002) has been shown in Figure 1. In general, a Question answering system consists of the following modules:

- **Question-Analysis Module:** Identifies the question type, answer type, topic and focus of the question and transforms question into a query.

- **Document-Processing Module:** Identifies a set of documents that are relevant to user’s query. Also, it filters the paragraphs that are likely to contain answer and rank them.

- **Answer-processing Module:** Extracts answers from ranked text segments obtained from the above module. The extracted answers are then ranked. The user is thus provided with a list of answers sorted in order of their relevance.

The QA systems that have been introduced in the literature collected the pages from the Web. However, it has been found that the information contained in web pages is not of high quality and also, the major portion of the web pages contain information that is not of much importance to the user. So, the web is a huge repository of information but is not likely to contain the information that is focussed. Using the general web pages for question answering doesn’t guarantee accurate answers. For this purpose, a system needs to be developed that must be able to collect information from some other sources that are likely to be containing focussed information. One of the richest sources of information is Blogs (Weblogs). Blog, in Cohen et al. (2006) is a personal journal maintained on the Web. Blogs are used for writing on a variety of topics. Blog is an important...
Related Content

Service-Oriented Architectures for Context-Aware Information Retrieval and Access
[www.igi-global.com/chapter/service-oriented-architectures-context-aware/28075?camid=4v1a](www.igi-global.com/chapter/service-oriented-architectures-context-aware/28075?camid=4v1a)

Sentiment Analysis Using Cuckoo Search for Optimized Feature Selection on Kaggle Tweets
[www.igi-global.com/article/sentiment-analysis-using-cuckoo-search-for-optimized-feature-selection-on-kaggle-tweets/217480?camid=4v1a](www.igi-global.com/article/sentiment-analysis-using-cuckoo-search-for-optimized-feature-selection-on-kaggle-tweets/217480?camid=4v1a)

A Collaborative Situational Method Engineering Approach for Requirement Gathering: A Re-Defined View
[www.igi-global.com/article/a-collaborative-situational-method-engineering-approach-for-requirement-gathering/193246?camid=4v1a](www.igi-global.com/article/a-collaborative-situational-method-engineering-approach-for-requirement-gathering/193246?camid=4v1a)
Speech Emotion Analysis of Different Age Groups Using Clustering Techniques
www.igi-global.com/article/speech-emotion-analysis-of-different-age-groups-using-clustering-techniques/193250?camid=4v1a