

ENHANCED INFORMATION RETRIEVAL MODEL USING CONTEXT BASED QUERYING

R.Sanmuga priya, D.Prabakar, S.Karthik

SNS College of Technology, Coimbatore, India

*Corresponding author: Email: priyarsns@gmail.com

ABSTRACT

Traditionally people use libraries for retrieving the information. In recent times Internet is used for this purpose. Search engines serve as a tool for knowledge hunting from the Internet. Nowadays, these search engines are prone to improve for the following reasons: First the available information is ever growing which leads to vast information and hence prone to contain duplicates also. Secondly, when the query is given by the user to the search engine, it returns a huge number of links from where the user needs to explore and extract the required information. This will lead to waste of time and frustration. To address these problems the proposed system automates the process of extracting useful data from a large pool of search result documents given by search engine as a result of search process on the World Wide Web. When a user initiates a search on the Web, the standard search engines help to find the web documents relevant to it. Our system automatically explores the contents of these search result documents, removes noise, preprocesses the data and extracts the most valuable relevant data. To achieve this rule based algorithms have been developed. This algorithm not only simplifies the search process but also provides just in time information. This is achieved by extracting information based on the context of search. This results in minimal seek time, non-duplicate context based information. The novelty of this system is, it attempts to answer the questions of the user based on the facts extracted from the search result documents.

INTRODUCTION

With the abundant availability of resources and information on the World Wide Web, access to its content is also increasing tremendously. Due to the exponential growth of information on Web and the prevalence of multiple sources with similar content, information retrieval on the Web has become a challenging task. Search engines are improving continuously to achieve the goal of effective search with minimal seek time. But still there is always a discontent by the users due to the volume of search results given by the search engines. The users are in want of most relevant information, concise and compact with minimal retrieval time. Our proposed system attempts to address this issue by obtaining results of search engines, removing the spurious data and providing the facts as per the context of the information seeker. The system is built on a rule based natural language processor, that is capable of analyzing and triggering the rules based on the context of questions posed by the user. This system is dynamic and domain independent. Our experimental results have shown significant improvement in eliminating the spurious and duplicate facts and providing answers to the queries with a greater degree of relevance.

RELATED WORK

Over the recent years, web information retrieval systems find and rank documents based on maximizing relevance to user query. The users of the web are from Heterogeneous background and the search processes on the web are very complicated. One is that the user himself may not be fully aware of the topic on which the search is initiated which leads to improper query formulation. The other problem is the volume and depth of information content provided by the search engines and web pages. Though many works focus on query reformulation and re-ranking of search results, the process of providing the right quantity of the required information is still an unexplored area. Recent times the techniques of text document summarization and knowledge extraction have been employed to overcome these problems. Dongfeng Cai et al, proposed a combination method for question classification based on the patterns and semantic resource. The system mainly focused on the interrogative words, question focus words. Gautam et al, proposed a system which used the relation between the nouns and the verbs in a sentence. The subject, verb, object of the sentences are identified as a key information in processing phase. The natural language processing method is used to parse and extract the information from the sentences in the document. Then semantically related keyword search in the extracted key information is performed to retrieve the document. Hoeber et al illustrated the benefits of supporting flexible interaction within a single unified web search interface. The information assists searchers in their web search tasks and the system supports user interaction during the query refinement process and search result exploration process. Jinzhong et al, proposed the question answering system was based on Natural language processing and introduced application theory and framework which were built by question semantic representation and ontology. In the works of Liu et al indicated a user interactive question answering system based on semantic pattern and answer analysis were done to improve correctness of extracted answers. Min-Kyoung Kim et al proposed an approach that provided the candidate questions for the users to select and also used the sentences within the documents as a source for providing the answers. They prepared a complete database for question/answering with the facts and events to analyze the whole

documents. Qinglin Guo et al proposed a high power software system based on internet. It is an interrelated technology based on natural language understanding with the knowledge base and corpus, word segmentation, tagging of text and grammatical analysis of sentences. They focused on knowledge information based on semantic network in question answering system, Syntax parse model and construction of question answering system. Vazquez-Reyes et al proposed that question answering systems extract answers rather than retrieval of relevant documents. The data collection analysis system suggested the appropriate texts in which reasoning and explanation that constitute an answer to a "why" question is capable of extracting from the source text. The implementation of extracting candidate answers from source text uses an approach which combines lexical overlapping and lexical semantic relatedness for ranking possible answers to causal questions. Wu et al, proposed the goal of the system such that they use the top results obtained from a search engine to extract and present correct answers. This limited the focus on identifying the correct answers to the questions which require deep natural language understanding. This involves a sequence of actions to answer "How do I" questions. To tackle the top documents retrieved for a set of queries sampling techniques are used. Yinli Wang et al introduced an intelligent question answering system which provided fast answers for frequently asked questions by what sentence similarity computation.

PROPOSED RETRIEVAL MODEL

Information explosion on the web is proliferating at an unprecedented speed that it is becoming difficult for the user to digest the information. Web is a collection of diverse information in terms of context, content, format and quality. However this diversity, as good as it is, often brings challenges for users in their web information seeking activities. Standard search engines on the Internet are able to answer Natural Language questions with long answers. They do not have the ability to identify the exact short answers to the questions. They can, at best, guess that the answer may be located in the sentence, but is not able to clearly identify the real answer. It has improved to search and list down the documents that are more relevant to the user search query. Lot of improvements on query formulation has helped in triggering better search process. But still the task of exploring the contents of the search result documents is mostly the task of the user. It is always the responsibility of the users to find their interested information from the returned documents which is a large junk of raw data. The objective of the system is to accept the search query from the user and automatically explore the contents available on the Web based on the search results obtained from the search engines. The obtained search contents are further preprocessed and spurious and duplicate information are removed. Thus extracted facts are presented to the user. Moreover natural language processing techniques are applied for question analysis and the system is modified as a question answering system based on context based information retrieval. The architecture diagram is as given in figure 4.1. The user gives the search query in the text box available in the user interface. The question answering system is integrated to an external search engine and using this query, it initiates a search process through the interface of the search engine (in our case Google). The top n search links are retrieved and stored in the search link repository. The answer generator retrieves the web document content of each search result link stored in the search result link repository and they pass through the Answer Generator sub-module. The Answer Generator analyzes the semantics of the question and uses the natural language processing rules and keyword selection algorithm to return the processed facts. These facts are ranked based on the relevance to the search query and then given to the user.

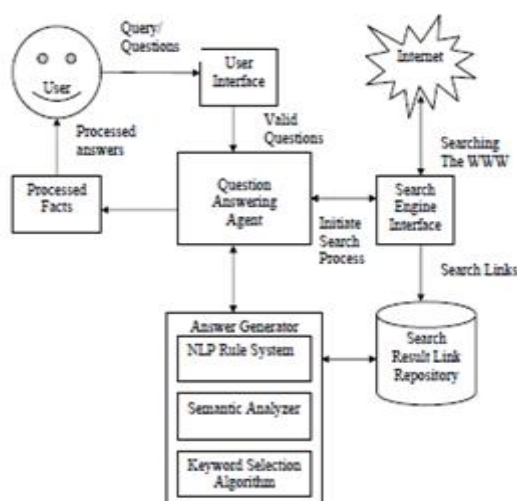


Figure.3.1. Proposed Information retrieval Model

IMPLEMENTATION

Our main task is to automate the knowledge extraction process on the web. This automation is achieved by presenting to the user only the relevant data in the form of facts. Each web page exists in HTML format, containing both the tags part and the text part. Before applying the algorithm, the web page must be detagged to extract the text alone from it. The extracted text must be preprocessed to generate the relevant facts.

Search Process Initiation

The search query given by the user to the user interface. The search links relevant to the query are obtained by the search engine and are stored in the search result link repository. The documents can then be extracted from these links stored in the repository.

Document Extraction

Once the search links are extracted, the documents of those search links alone are extracted from the web by means of web spiders.

Preprocessing

The web documents extracted by the web spiders are generally in HTML format. In addition it may contain certain redundant and irrelevant data. Hence preprocessing must be done to make the data more concise and informative. The various steps in preprocessing are listed as follows:

- Stop word removal
- Detagging
- Noise elimination
- Stemming

Stop word Removal

Stop words are common words which play very little or no role in the processing of Natural Language data. Hence they ought to be filtered out. In this step all the stop words are removed from the question.

Detagging

The first step in preprocessing is detagging. This is the process of removing all the HTML tags from the web page and converting the HTML document content into text content.

Noise Elimination

It is important to make it sure that the document used in the Knowledge Extraction process is noise free. Hence noise is removed from the document by applying a set of rules to the detagged file. Some of the rules for noise elimination are as follows:

- The sentence should have more than four words after the detagging process.
- The sentence should not have any special characters.
- The sentence should not start with words such as Where, What, Why, How.
- The sentence should not have words such as copyright.
- The sentence should not be a URL.
- The first letter of the first word of the sentence should be in uppercase.

Context Based Fact Retrieval

The content of each web document is grouped as collection of valid facts. After the elimination of noisy facts, the next step is to pool in all the facts from various web documents. To carry out this, the process of Stemming is done by comparing all the words to their root words. For example, words like 'networks', 'networking', and 'networked' will be compared to its root word 'network'.

Step 1: Break the noise removed file into sentences.
Step 2: Compare each word in the file to their corresponding root word of the keyword by applying the Stemming algorithm.
Step 3: If the sentence contains the keyword then accept the sentence as the fact.
Step 4: Store these preprocessed facts into a new file.

Question Based Fact Retrieval

Once the valid facts are retrieved based on the context, the next step is to select in all the facts which answers to the question/query given by the user. These factual answers are retrieved from the context based fact retrieval process. To carry out this process, a set of patterns are considered:

Why ⇒ REASON

When ⇒ DATE

What ⇒ DESCRIPTION

How much/many ⇒ QUANTITY

How ⇒ DESCRIPTION

RESULTS

Sample Question 1: **What is a network?**

Table 2 Comparative Search for Top ten results

Link Google	Google	System	Received / actual %
1	902	756	83.81375
2	3043	1670	54.88005
3	174	89	51.14943
4	921	588	63.84365
5	465	328	70.53763
6	148	83	56.08108
7	315	291	92.38095
8	561	372	66.31016
9	589	286	48.55688
10	319	183	57.36677

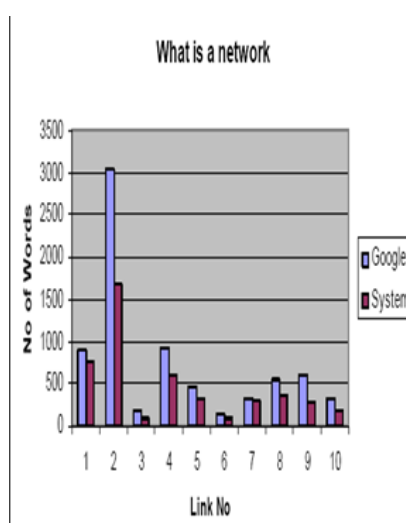


Figure 5.1. Comparison of volume of content in top ten search results

CONCLUSION

Context based question answering system has been developed to help the web users with an enhanced search process. It complements the search engine process by exploring the contents of the web search link results. The link contents are preprocessed and based on natural language rule based system, only the valid facts are retrieved. The queries are matched with their context relevant facts and the results are given to the user. It eases the burden of exploring the contents of web sites to acquire the required information with minimal seek time. The system can be tagged with multiple search engines for hybrid information extraction and it presents the user with the most appropriate facts to the given query.

FUTURE ENHANCEMENTS

The extraction of web documents have been done from search results of one search engine. This can be upgraded by having hybrid search results by combining the search results of more than one search engine. The question answering system can be upgraded to accommodate different classifications of questions and also to adhere to semantic analysis of the questions and the retrieved facts. The system can be upgraded to possess intelligence by which the user profiles can be tracked and the facts that interest a particular user can be given using push technology. This system can be used in extracting facts from web documents and filtering out the noisy and redundant facts. It obtains the facts and gives context based facts, question based facts and all the facts to the user. Since it is not domain based, it caters to a wide spectrum of users. This can also be enhanced in various aspects. Some of the future enhancements are listed below.

•Effective search results

The extraction of web documents have been done from search results of one search engine. This can be upgraded by having hybrid search results by combining the search results of more than one search engine.

•Semantic answering System

The question answering system can be upgraded to accommodate different classifications of questions and also to adhere to semantic analysis of the questions and the retrieved facts.

•**Knowledge based answering system**

The system can be upgraded to possess intelligence by which the user profiles can be tracked and the facts that interest a particular user can be given using push technology.

REFERENCES

- Dongfeng Cai, Yu Bai, Yanju Dong and Lei Liu, "Chinese Question Classification Using Combination Approach", Third International Conference, 29-31 Oct. 2007.
- Gautam, Miyoung D. and Cho Pankoo Kim, "Document Retrieval Based on Key Information of Sentence", ICACT 2008 10th International Conference, 17-20 Feb. 2008.
- Hoeber, Orland Brooks, Michael Schroeder, Daniel Yang and Xue Dong, "The HotMap.com: Enabling Flexible Interaction in Next-Generation Web Search Interfaces", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 9-12 Dec. 2008.
- Jinzhong Jia, Keliang Fu and Jibin, "Research of Automatic Question Answering System in Network Teaching", International Conference, 18-21 Nov. 2008.
- Liu Xiaoli, Wu Guoqing, Jiang Min, Yang Min and Wang Weiming, "Software architecture for a pattern based Question Answering system", ACIS International Conference, 20-22 Aug. 2007.
- Min-Kyoung Kim and Han-Joon Kim, "Design of Question Answering System with Automated Question Generation", NCM '08 Fourth International Conference, 2-4 Sept. 2008.
- Qinglin, Guo, Kehe, Wu, Wei and Li, "The Research and Realization about Question Answer System based on Natural Language Processing", ICICIC'07 Second International Conference, 5-7 Sept. 2007.
- Vazquez-Reyes, Sodel Black and William J, "Evaluating Causal Questions for Question Answering", ENC '08. Mexican International Conference, 10 Oct. 2008.
- Wu, Lei K. and Yu Cutler, "Towards Answering How do I Questions Using Classification", AINAW'07 21st International Conference, 21-23 May 2007.
- Yinli Wang and Guanglai Gao, "Research and Implementation of Intelligent Question Answering System in a Restricted Domain", CCPR '08 Chinese Conference, 22-24 Oct. 2008.