World Engineering & Applied Sciences Journal 6 (2): 132-135, 2015
ISSN 2079-2204

DOI: 10.5829/idosi.weasj.2015.6.2.22168

Enhanced Information Retrieval Modelusing Context Based Querying

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

SNS College of Technology, Coimbatore, India

Abstract: Traditionally people use libraries for retrieving the information. In recent timesInternet is used for this purpose. Search engines serves as a tool for knowledgehunting 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.

Key words: Depth of information · Power software system · Significant improvement · Multiple sources · Natural language

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.* [1], proposed a

Corresponding Author: R. Sanmuga priya, SNS College of Technology, Coimbatore, India.

combination method for questionclassification based on the patterns and semantic resource. The systemmainly focused on the interrogative words, question focus words. Gautam etal [2], proposed a system which used the relation between the nouns and theverbs in a sentence. The subject, verb, object of the sentences are identifiedas a key information in processing phase. The natural language processingmethod 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* [3] 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* [4], 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*. [5] 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*. [6] proposed an approach that provided the candidate questions for the users toselect and also used the sentences within the documents as a source forproviding the answers. They prepared a complete database for question/answering with the facts and events to analyze the whole documents. Qinglin Guo *et al*. [7] 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 onknowledge information based on semantic network in question answeringsystem, Syntax parse model and construction of question answering system.Vazquez-Reyes *et al*. [8] proposed that question answering systems extract answers rather than retrieval of relevant documents. The datacollection analysis system suggested the appropriate texts in whichreasoning and explanation that constitute an answer to a "why" question iscapable of extracting from the source text. The implementation of extractingcandidate answers from source text uses an approach which combineslexical overlapping and lexical semantic relatedness for ranking possibleanswers to causal questions. Wu *et al* [9], proposed the goal of the systemsuch that they use the top results obtained from a search engine to extractand present correct answers. This limited the focus on identifying the correct answers to the questions which require deep natural languageunderstanding. This involves a sequence of actions to answer "How do I"questions. To tackle the topdocuments retrieved for a set of queriessampling techniques are used. Yinli Wang *et al*. [10] introduced an intelligentquestion answering system which provided fast answers for frequently askedquestions by what sentence similarity computation.

Proposed Retrieval Model: Information explosion on the web is proliferating at an unprecedentedspeed that it is becoming difficult for the user to digest the information. Webis a collection of diverse information in terms of context, content, format andquality. However this diversity, as good as it is, often brings challenges forusers in their web information seeking activities. Standard search engines onthe Internet are able to answer Natural Language questions with longanswers**.** They do not have the ability to identify the exact short answers tothe questions. They can, at best, guess that the answer may be located in thesentence, but is not able to clearly identify the real answer. It has improvedto search and list down the documents that are more relevant to the user search query. Lot of improvements on query formulation has helped intriggering better search process. But still the task of exploring the contents ofthe search result documents is mostly the task of the user. It is always the responsibility of the users to find their interested information from thereturned documents which is a large junk of raw data. The objective of thesystem is to accept the search query from the user and automatically explorethe contents available on the Web based on the search results obtained fromthe search engines. The obtained search contents are further preprocessedand spurious and duplicate information are removed. Thus extracted factsare presented to the user. Moreover natural language processing techniquesare applied for question analysis and the system is modified as a questionanswering system based on context based information retrievalThe architecture diagram is as given in Figure 4.1. The user gives thesearch query in the text box available in the user interface. The questionanswering system is integrated to an external search engine and using thisquery, 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 thesearch link repository. The answer generator retrieves the web documentcontent of each search result links stored in the search result link repositoryand they pass through the Answer Generator sub-module. The Answer.

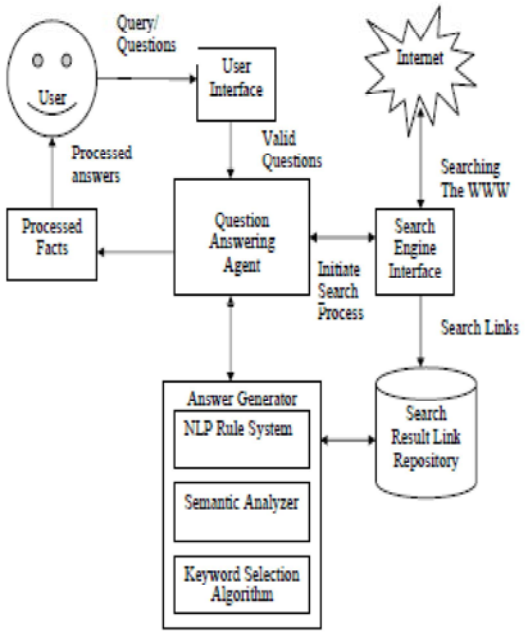

Fig. 3.1: Proposed Information retrieval Model

Generator analyzes the semantics of the question and uses the naturallanguage processing rules and keyword selection algorithm to returns theprocessed facts. These facts are ranked based on the relevance to the searchquery and then given to the user.

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 partand the text part. Before applying the algorithm, the web page must be detagged to extract the text alone from it. The extracted text must bepreprocessed to generate the relevant facts.

Search Process Initiation: The search query given by the user to the user interface. The searchlinks relevant to the query are obtained by the search engine and are storedin the search result link repository. The documents can then be extractedfrom these links stored in the repository.

Document Extraction: Once the search links are extracted, the documents of those search linksalone are extracted from the web by means of web spiders.

Preprocessing: The web documents extracted by the web spiders are generally inHTML format. In addition it may contain certain redundant and irrelevantdata. Hence preprocessing must be done to make the data more concise andinformative. 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 theprocessing of Natural Language data. Hence they ought to be filtered out. Inthis step all the stop words are removed from the question.

Detagging: The first step in preprocessing is detagging. This is the process ofremoving all the HTML tags from the web page and converting the HTMLdocument content into text content.

Noise Elimination: It is important to make it sure that the document used in theKnowledge Extraction process is noise free. Hence noise is removed fromthe 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 detaggingprocess.
- 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 validfacts. After the elimination of noisy facts, the next step is to pool in all thefacts from various web documents. To carry out this, the process of Stemming is done by comparing all the words to their root words. Forexample, words like ‘networks’, ‘networking’ and ‘networked’ will becompared 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 isto select in all the facts which answers to the question/query given by theuser. These factual answers are retrieved from the context based factretrieval process. To carry out this process, a set of patterns are considered:

Why □ REASON
When □ DATE
What □ DESCRIPTION
How much/many □ QUANTITY
How □ DESCRIPTION

Result Analysis:

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

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

Fig. 5.2: Comparative Search for Top ten results

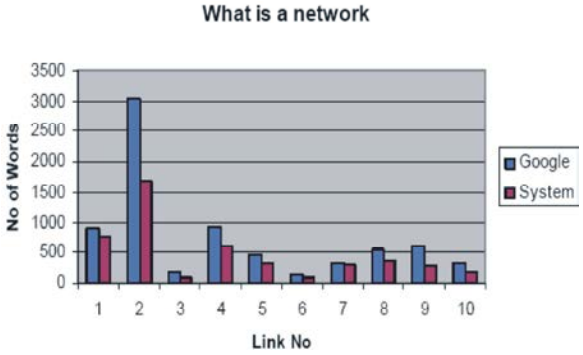

Fig. 5.1: Comparison of volume of content in top ten search results

CONCLUSION

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

REFERENCES

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