An Advance Way to Retrieve Information from Web by Natural Language Interface

Dr. Mayura Mathankar

***Abstract*— Information is one of the most important part of our today's daily life. There are large numbers of things that can be done using the information and so it is very important. We can say that with the progress in the information technology, we are progressing in every sphere of life as it not only makes our tasks easier but also saves a lot of time. To access any information, internet is used for different purposes depending upon the requirement. People mostly rely on internet for acquiring desired information. Search engine plays vital role in retrieving the required information in fast way.**

Many existing state-of-the-art search engines face challenges of effective information retrieval. This research paper deals with the design of Interface for Web-based information Retrieval with the help of Natural Language. In this proposed model, natural language query is given to the developed search engine to understand its meaning and depending on its meaning the most relevant documents are retrieved from the World Wide Web.

The proposed algorithm is based on Statistical Machine Learning. Natural Language text is converted to more formal representation such as first-order logic structure because it is easier for computer programs to manipulate. The aim of this research work is to provide interactive interface to the search engine to enhance its effectiveness for information retrieval.

***Index Terms*—Natural Language, Algorithm, Web-based information.**

I. INTRODUCTION

From computers and smartphones to watches and eyewear, almost anything can now be connected to what is essentially one huge global network, which is constantly evolving further. The internet has revolutionised communication and networking to the point where few people in the world cannot be reached in a few seconds at the click of a button. Gone are the days of sending letters and waiting days for a reply, being the only person in a specific area with a specific interest or the inability to speak to people face-to-face without being in proximity of them, replaced by the instant contact method of email. People are increasingly accustomed to have resource to computer programs to do the computations, to search needed information and even to make decisions. Therefore, an obliging and clever interface will greatly reduce user's work and increase the efficiency of the computer programs.

1.1 Intelligent Interfaces

Intelligent interface agents are computer programs that employ artificial intelligence techniques in order to provide assistance to a user dealing with a particular computer application. Interface agent must have some kind of intelligence, such as knowledge acquisition, autonomy and collaboration. The architecture of intelligent interface includes rules, frame descriptors, discrimination networks, inference engine, associative memory, matching and autonomous agents.

Intelligent user interfaces have been proposed as a means to overcome some of the problems that direct manipulation interfaces cannot handle, such as information overflow problems; providing help as how to use complex systems; or real time problems. Intelligent user interface is also being proposed as a means to make systems individualized or personalized. Interface agents are the prevailing development towards intelligent user interfaces.

An intelligent interface cooperates with the user in performing its tasks, working as a personal user assistant. The interface agent is pro-active taking the initiative and not passive. Usually, an intelligent interface is not the interface between the user and the application. Instead, it observes the interactions between the user and the program learns with it and interacts both with user and program.

1.2 Web-based information retrieval

The Internet and the Web offer new opportunities and challenges to information retrieval researchers. With the information explosion and never ending increase of web pages as well as digital data, it is very hard to retrieval useful and reliable information from the Web. Materials from millions of web pages from organizations, institutions and personnel have been made public electronically accessible to millions of interested users. The Web uses an addressing system called Uniform Resource Locators (URLs) to represent links to documents on web servers. These URLs provide location information. Like titles of books in traditional libraries, no one can remember all URLs on the Web. Web search engines allow us to locate the Internet resources through thousands of Web pages. It is almost impossible to get the right information as there is too much irrelevant and out dated information.

Information retrieval systems provide useful information. The Web can be viewed as a virtual library. Information retrieval is an important and major component of the Internet. Most of the current search engines are based on words, not the concepts. When searching for certain information or knowledge with a search engine, one can only uses a few key words to narrow down the search. The result of the search is tens or maybe hundreds of relevant and irrelevant links to various Web pages.

Dr. Mayura Mathankar, Associate Professor, Sydenham College of Commerce & Economics, Mumbai

Document representation, query formulation, and retrieval functions are fundamental issues in information retrieval study. Based on this classification, we have to design and implement an appropriate scheme to represent the contents of documents, a language to express user queries, and a retrieval function to search for relevant documents. Index terms play the connecting role between documents and user queries. A document is considered to be relevant to a query if the user submitting the query judges the document to be useful.

Different retrieval models have been developed, such as the Boolean, vector space, and probabilistic models, as well as recent linguistic and knowledge-based models. The first three models are often referred to as the exact match model; the latter as the best match models. Although they provide us formal and elegant formulations of information retrieval problems, they suffer from several shortcomings. The classical retrieval models are over-simplification of the real world retrieval problem.

II. METHODOLOGY

A natural language search engine would find targeted answers to questions. For example, when confronted with a question of the form ‘which Indian state has the highest income tax?', conventional search engines ignore the question and instead search on the keywords 'state, income and tax'. Natural language search, on the other hand, attempts to use natural language processing to understand the nature of the question and then to search and return a subset of the web that contains the answer to the question. If it works, results would have a higher relevance than results from a keyword search engine. In interface design natural language interfaces are sought after for their speed and ease of use, but most suffer the challenges to understanding wide varieties of ambiguous input.

It is important to note that text interfaces are 'natural' to varying degrees, and that many formal (un-natural) programming languages incorporate idioms of natural human language. Likewise, a traditional keyword search engine could be described as a 'shallow' Natural language user interface. If search engine is supplemented by an intelligent interface, it may help users to get relevant information effectively and efficiently which can result in optimum utilization of e-resources by potential users.

III. INTELLIGENT NATURAL LANGUAGE INTERFACE

Many intelligent interfaces exhibit mainly three features such ie Knowledge Acquisition, Autonomy and Collaboration. The Intelligent Natural Language Interface for Web – based information retrieval is comprised of following components:

1 Knowledge Acquisition :

It is the process of absorbing and storing new information in memory, the success of which is often gauged by how well the information can later be remembered.

2 Knowledge representation

For knowledge representation First- Order Predicate Logic will be used. First-order logic (like natural language) assumes the world contains:

– Objects: people, city, numbers, colors, football etc.
– Relations: red, round, prime, brother of, bigger than, part of etc
– Functions: father of, best friend, one more than, plus etc

The syntax of FOPL consists of following elements

- Constants→Krishna, 2, NUS,...
- Predicates→Brother, >,...
- Functions→Sqrt, Left, Upper
- Variables→x, y, a, b,...
- Connectives→ $\neg, \Rightarrow, \wedge, \vee, \Leftrightarrow$

3 Rules

Complex deductive arguments can be judged valid or invalid based on whether or not the steps in that argument follow the basic rules of inference. These rules of inference are all relatively simple, although when presented in formal terms they can look overly complex.

4 Inference Engine

An inference engine is a software system that is designed to draw conclusions by analysing problems in light of a database of expert knowledge it draws upon. It reaches logical outcomes based on the premises the data establishes. Sometimes inference engines are also capable of going beyond strict logical processing, and utilize probability calculations to reach conclusions that the knowledge database doesn't strictly support, but instead merely implies or hints at.

5 Semantic Parser

Semantic parser checks syntax of the text, i.e. whether the text is grammatically correct or not. It further analyses the input text for understanding its meaning. The meaning of any natural language sentence depends on the way of parsing the text and its context. For example, consider a sentence, “When I sing well children feel sick”. This sentence implies two meanings based on how we parse it.

Ex.1 : When I sing well, children feel sick

Ex. 2: When I sing, well children feel sick

6 Search Engine

A web search engine is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

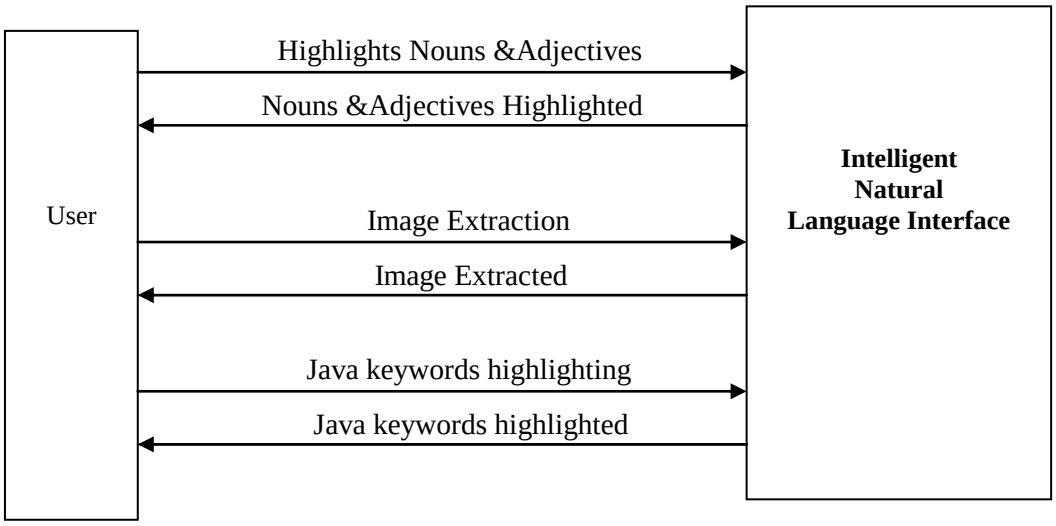

Figure 1: DFD for Intelligent Natural Language Interface

IV. RESULTS

The proposed system is developed and implemented in JAVA. Rigorous experimentation was carried out. The results are tested for various sample queries.

Precision and recall

	Average Precision	Average Recall
Without interface	0.64	0.72
With interface	0.81	0.76

V. CONCLUSION

In this paper we have presented the design of Natural Language Interface for Web-based information Retrieval. In this model, natural language query is analysed for understanding its meaning and depending on its meaning the most relevant documents are retrieved from the World Wide Web. Digital document libraries have become an increasingly important means of storing information within organizations. Automatic understanding of documents will certainly improve the utilization of e-resources among the potential users.The aim of this work is provide interactive interface to the search engine to enhance its effectiveness for information retrieval.

REFERENCES

[1] Han, H., "Extracting news from server side databases by query interfaces" (2014) Journal of Computer Information Systems, 54 (2), pp. 57-65

[2] Cobos, C., Muñoz-Collazos, H., Urbano-Muñoz, R., Mendoza, M., León, E., Herrera-Viedma, E., Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion (2014) Information Sciences, 281, pp. 248-264

[3] Sleiman, H.A., Corchuelo, R., A class of neural-network-based transducers for web information extraction (2014) Neurocomputing, 135, pp. 61-68.

[4] A.K.Rabiah, T.M.T. Sembok, B.Z. Halimah, *"Improvement of document understanding ability through the notion of answer literal expansion in logical-linguistic approach"*, WSEAS Transactions on Information Science and Applications, Vol. 6, Issue 6, June 2009

[5] Marco Aiello, ChristofMonz, Leon Tororan, Marcel Worring, "*Document Understanding for a Broad Class of Documents*", International Journal on Document Analysis and Recognition.

[6] Javier Albusac, David Vallejo, and J.J. Castro-Schez, Paolo Remagnino, Carlos Glez Morcillo and Luis Jimenez, **"***Monitoring Complex Environments Using a Knowledge- Driven Approach Based on Intelligent Agents",* IEEE Journal of Intelligent Systems,May- June 2010, Vol. 25, no. 3, pp : 24 - 31