Improving Web Search with the EWEBSEARCH Model

Abur M. M.
Iya Abubakar Computer Centre, Faculty of Science,
Ahmadu Bello University, Nigeria

Hammawa M. B.
Mathematics Department, Faculty of Science,
Ahmadu Bello University, Nigeria

Adewale S. O.
Mathematics Department, Faculty of Science,
Ahmadu Bello University, Nigeria

Soroyewun M. B.
Iya Abubakar Computer Centre, Faculty of Science,
Ahmadu Bello University, Nigeria

ABSTRACT

Traditional web which is the largest information database lacks semantic and as a result the information available in the web is only human understandable, not by machine. With the rapid increase in the amount of information on networks, search engine has become the infrastructure for people gaining access to Web information, and is the second largest Internet application besides e-mail. However, search engine returns a huge number of results, and the relevance between results and user queries is also different. There are lots of search engines available today, but the way to retrieve meaningful information is difficult. To overcome this problem in search engines to retrieve meaningful information intelligently or smartly, Semantic Web technology has played a major role. In the light of this, our paper, proposes an algorithm, architecture for the semantic web based search engine named EWEBSEARCH model, powered by XML meta-tags (which ensures machine understandability) to improve web search. The EWEBSEARCH model provides a simple interface to capture user's queries (keywords), then the search or query engine processes the queries from the repository (database) using the search engine algorithm, interpreting the queries, retrieving and providing appropriate ranking of results in order to satisfy users queries. Query answers are ranked using extended information-retrieval techniques, are generated in an order of ranking and implementation of the model.

General Terms

Architecture, EWEBSEARCH model, Algorithms et. al.

Keywords

Database, EWEBSEARCH model, Search engine, Semantic Web, XML meta-tags.

1. INTRODUCTION

World Wide Web has changed the way people communicate with each other and the way business is conducted. It lies at the heart of a revolution that is currently transforming the developed world toward a knowledge economy and, more broadly speaking, to a knowledge society. This development has also changed the way we think of computers. Originally they were used for computing numerical calculations. Currently their predominant use is for information processing, typical applications being database systems, text processing, and games. At present there is a transition of focus toward the view of computers as entry points to the information highways. Most of today's Web content is suitable for human consumption. Even Web content that is generated automatically from databases is usually presented without the original structural information found in databases (Antoniou et al, 2008). Typical uses of the Web today involve people's seeking and making use of information, searching for and getting in touch with other people, reviewing catalogs of online stores and ordering products by filling out forms, and viewing adult material. These activities are not particularly well supported by software

2. RELATED WORKS

All material Google, Yahoo and Bing have been out there which handles the queries after processing the keywords, which makes them keyword based search engine. They only search information given on the web page. Recently, some research groups start delivering results from their semantic-based search engines; however most of them are in their initial stages.

Evri is a semantic search engine. Evri seeks to build a "map of connections between people, places, and things on the Web" (Evri: About Us). With Evri, a user can view a result page for a single subject from a pre-existing list of popular subjects.

Hakia (Tümer et al, 2009) is a general purpose semantic search engine that search structured text like Wikipedia. Hakia calls itself a "meaning-based (semantic) search engine". They're trying to provide search results based on meaning match, rather than by the popularity of search terms. The presented news, Blogs, Credible, and galleries are processed by hakia's proprietary core semantic technology called QDEXing (Tümer et al, 2009). It can process any kind of digital artefact by its SemanticRank technology using third party API feeds. A single query by the user brings results from any repository including Web, News, Blogs, Video, Images, Hakia Galleries and also from Credible Sources For short queries the site displays results in categories, instead of a standard list as shown in current search engines. For longer queries, Hakia highlights relevant phrases or sentences. The results are somehow relevant and reliable but Hakia does not reveal it's inside technology. Hakia take the searched query and find the results in many categories for example from galleries, videos so it takes more time than the usual search engines in the retrieval of results (Tümer et al, 2009).

SenseBot (Sensebot et al, 2010) represents a new type of search engine that prepares a text summary in response to the user's search query. SenseBot extracts the most relevant results using Semantic Web technologies from the Web. It then summarizes the results together for the user as per topic. It uses text mining algorithms to parse (human readable) Web pages which lead to identification of key semantic concepts. The coherent summary is then performed from multi-documents that are retrieved. This summary itself becomes the main result of the search. Although the search results are still not relevant, this is because the summarized result may divert the results from actual demands of the user. The sources from which the results are coming are usually the news agencies so reliability is also somehow missing (Sensebot et al, 2010).

JEESSE- Journal of Energy & Enviroment Semantic Web Search Engine is a semantic search engine that allows the lecturers and other users form department of Energy and Environment of the Malaysian University in searching for related article using a single keyword. (Alhassan et al, 2011). Hence, it uses single keyword search and the searching is limited to the Energy and Environment only.

3. THE PROPOSED SYSTEM

Our proposed EWEBSEARCH model (Abur et al, 2012) tends to address the short comings of the existing semantic search engines, by providing a better User interface, Query engine, and database.

3.1 User Interface

The User interface usually called the query interface is the page that users see when they navigate to the search engine to enter a search term (ledford et al, 2008). Here, we overcome the limitation of current keyword-based semantic search engines by supporting a Google-like user interface which supports queries in terms of multiple keywords. Google-like User Interface Layer (see figure 1), which allows end users to specify queries in terms of keyword Google-like query interface, extends traditional keyword search languages by allowing the explicit specification of the queried subject or multiple keywords. The proposed user interface for the EWEBSEARCH model provides a simple, flexible, and powerful approach for specifying user queries.

3.2 Database

A database is an organized collection of data, today typically in digital form. The data are typically organized to model relevant aspects of reality (for example: articles, authors, researchers), in a way that supports processes requiring this information. (Example: looking for a particular author of an article). The term database is correctly applied to the data and their supporting data structures, and not to the database management system (DBMS). The proposed model just like every search engine is connected to a system of databases.

3.3 The Search or Query Engine

This is the program that processes the queries (word or phrase) collected by the User interface from users check and locate the keywords in a large data source or database and returns the results (Ledford et al, 2008). Our proposed EWEBSEARCH model is concerned with the backend activities in order to present a simple user interface. First, it overcomes the problem of knowledge overhead suffered in formal query fronted search engines and form-based semantic search engines, as it does not require end users to be familiar with any semantic data, or any special query language. Second, the query interface provides a more flexible way of specifying queries than the interface presented by form-based search engines. Indeed, it does not confine users to any pre-defined query subjects and values. Third, in contrast with current semantic-based keyword search engines which only accept one keyword as input, this query interface supports the specification of complex queries in the format of specifying multiple keywords. Finally, this query interface is simpler than question answering tools as the search engine does not need to spend time calculating which of the keywords are in a user's query. In this paper, we focus on the user queries in which there are one or more keywords involved, in order to better explain the distinctive features of our search engine. The proposed EWEBSEARCH model (Abur et al, 2012) processes the query entered by users (multiple keywords or phrase of words) using the search engine algorithm, interpreting the query and providing appropriate ranking of results in order to satisfy users search and does the retrieval answers to satisfy users query. And it returns semantically related document fragments that satisfy the user's query. Query answers are ranked using extended information-retrieval techniques and are generated in an order of ranking. We use the power of XML meta-tags deployed on the web page stored in the database to search the queried information. The XML pages consist of user defined tags. The metadata information of the pages is stored and extracted from the database.

4. THE ALGORITHM FOR THE PROPOSED MODEL

For the sake of this paper, we used probability distribution to do the PageRank. Probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page in which he is searching for and the search algorithm is presented below with an example.

Example: Given a user query q, there is a set of documents which contain exactly the relevant documents and no other documents, called the ideal answer set. The query is a process for specifying the properties of the answer set, but we don't know what these properties are. Therefore an effort has been made to guess a description of the answer set and retrieve an initial set of documents. Then the user inspects the top retrieved documents, looking for the relevant ones. The proposed EWEBSEARCH model uses this information to refine the description of the ideal answer set. By repeating this process, it is expected that the description of the ideal answer set will be improved. The description of ideal answer set is modelled in probabilistic terms. Given a user query q and a document dj, the probabilistic model tries to estimate the probability that the user will find the document dj relevant. The model assumes that this probability of relevance depends only on the query and the document representations. The ideal answer set is referred to as R and 1 maximizes the probability of relevance. Documents in the set R are predicted to be relevant.

The probabilistic ranking is computed:

$$sim(dj,q) = P(R \mid dj) / P(\neg R \mid dj)$$

This is the ratio of the probability that the document dj is relevant and the probability that it is not relevant. It reflects the odds of the document dj being relevant, and minimizes the probability of an erroneous judgment. Using Bayes rule (for two events A and B, the probability of A given B is

P(A|B) = P(B|A) P(A) / P(B))

We expand the formula:

$$\mathrm{Sim}\ (\mathrm{dj}, \mathrm{q}) = \frac{\mathrm{P}(\mathrm{dj} \mid \mathrm{R}).\mathrm{P}(\mathrm{R})}{\mathrm{P}(\mathrm{dj} \mid \neg\mathrm{R}).\mathrm{P}(\neg\mathrm{R})} \cong \frac{\mathrm{P}(\mathrm{dj} \mid \mathrm{R})}{\mathrm{P}(\mathrm{dj} \mid \neg\mathrm{R})}$$

P(dj | R) is the probability of randomly selecting the document dj from the set R of relevant documents. P(R) stands for the probability that a document randomly selected from the document collection is relevant. The meanings attached to P(dj | ¬R) and P(¬R) are analogous and complementary. P(R) and P(¬R) are the same for all the documents relative to the query. We replace the probability of each document by the product of the probabilities of the terms it contains. We assume the terms occur in a document independent of each other; this is a simplifying assumption that works well in practice, even if in reality terms are not independent, the presence of a term might trigger the presence of a closely related term. We obtain:

$$\mathrm{Sim}\ (\mathrm{dj}, \mathrm{q}) \cong \frac{\prod_{\mathrm{i}} \mathrm{P}(\mathrm{ki} \mid \mathrm{R}) \,.\, \prod_{\mathrm{i}} \mathrm{P}(\mathrm{ki} \mid\neg \mathrm{R})}{\prod_{\mathrm{i}} \mathrm{P}(\neg\mathrm{ki} \mid \mathrm{R}) \,.\, \prod_{\mathrm{i}} \mathrm{P}(\neg\mathrm{ki} \mid\neg \mathrm{R})}$$

Where P(ki | R) is probability that the index term ki is present in a document randomly selected from the set R of relevant documents and P(¬ki | R) is the probability that ki is not present . The probabilities for ¬R have analogous meanings. Taking logarithms and ignoring factors that are constant for all the documents in the context of the same query we obtain:

$$\mathrm{Sim}\ (\mathrm{dj}, \mathrm{q}) \cong \sum \mathrm{wiq}\ \mathrm{wij} \left(\log \frac{\mathrm{P}(\mathrm{ki} \mid \mathrm{R})}{\mathrm{P}(\neg\mathrm{ki} \mid\neg \mathrm{R})} + \log \frac{\mathrm{P}(\mathrm{ki} \mid \neg\mathrm{R})}{\mathrm{P}(\neg\mathrm{ki} \mid\neg \mathrm{R})} \right)$$

Where w are binary weights, 1 if the index term is in the document or in the query, 0 if not.

P(¬ki | R)= 1 – P(ki | R) and P(¬ki | ¬R) = 1 – P(ki | ¬R).

The probabilities left to estimate are: P(ki | R) and P(ki | ¬R). They can have initial guesses:

P(ki | R) = 0.5 and P(ki | ¬R) = dfi / N, where dfi is the number of documents that contain ki.

This initial guess is used to retrieve an initial set of document V, from which the subset Vi contains the index term ki. The estimates are re-evaluated:

P(ki | R) = Vi / V and P(ki | ¬R) = (dfi – Vi) / (N – V)

This process can be repeated recursively.

Step1: User's queries (keyword (s)) q, are supplied in the graphical user interface.

Step2: The search engine then searches from the database, processes and interprets users query using the above probabilistic formula.

Step 3: The user query is then retrieved from the database as XML metadata.

Step 4: the results are then checked and related documents are ranked according to users queries using the probabilistic ranking formula:

$$\mathrm{Sim}\ (\mathrm{dj}, \mathrm{q}) = \frac{\mathrm{P}(\mathrm{dj} \mid \mathrm{R}).\mathrm{P}(\mathrm{R})}{\mathrm{P}(\mathrm{dj} \mid \neg\mathrm{R}).\mathrm{P}(\neg\mathrm{R})} \cong \frac{\mathrm{P}(\mathrm{dj} \mid \mathrm{R})}{\mathrm{P}(\mathrm{dj} \mid \neg\mathrm{R})}$$

Step 5: The related documents are then displayed in textual form using this probabilistic formula P(ki | R) = Vi / V and P(ki | ¬R) = (dfi – Vi) / (N – V)

Step 6: Users can also view the XML metadata of the displayed results.

5. USE CASE MODEL FOR THE PROPOSED EWEBSEARCH

In the design process phase, the use-case model has been used to design the process for the proposed WEBSEARCH model (Abur et al, 2012). The Use model is based on the concept of 'community'. A community refers to a set of people or computers that share the same characteristics and behaviour. Generally, there are two types of communities identified here: (Users and the Search engine). The propose model for improving Web search using Semantic Web technology consist of two actors (Users and the EWEBSEARCH engine), processes and four main events includes: (crawling and indexing, Response or Retrieval and ranking). Fig 1 shows the user case diagram of the proposed model. So users can search for any article using single or multiple keywords supplied to Graphical user interface (GUI). The Search engine then searches the keywords based on information contained in the metadata. We use the power of XML meta-tags deployed on the web page stored in the database to search the queried information. The XML pages consist of user defined tags. The metadata information of the pages is stored and extracted from the database. The proposed EWEBSEARCH (Abur et al, 2012) returns semantically related document fragments that satisfy the user's query. Query answers are ranked using extended information-retrieval techniques and are generated in an order of the ranking. Users can view their search result in textual form or semantically using XML metadata.

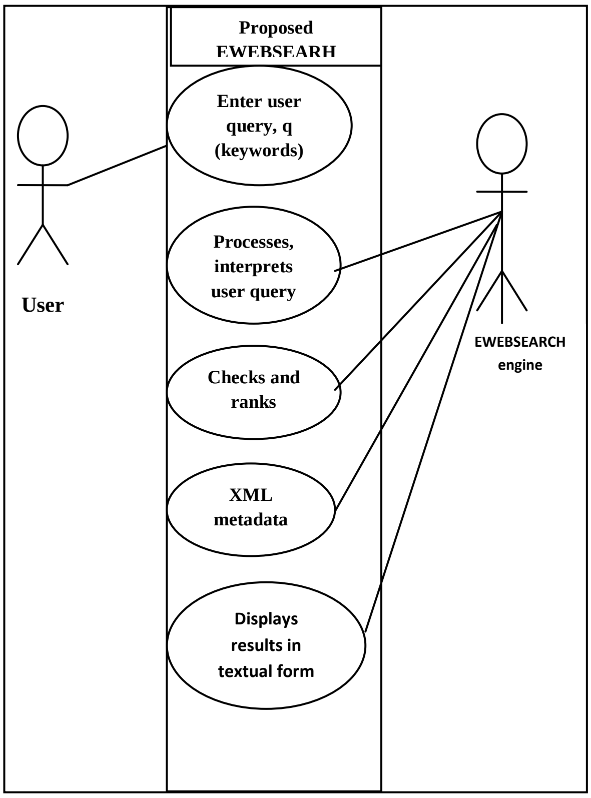
Proposed
EWEBSEARH
Enter user query, q (keywords)
Processes, interprets user query
Checks and ranks
XML metadata
Displays results in textual form
User
EWEBSEARCH engine

Fig 1: Shows the Use CASE diagram for the proposed EWEBSEARCH model.

6. IMPLEMENTATION OF THE EWEBSEARCH MODEL

In order to develop the prototype for improving Web search using semantic Web technology, we used Apache as a Web Server, MYSQL as a relational database and PHP scripting language to generate the Extensible Markup language (XML) metadata, the power of XML meta-tags deployed on the web page stored in the database to search the queried information. Editix- XML editor was also used to test for XML metadata.

Lines of codes segment in Fig2 below shows how searching is done with our EWEBSEARCH model:

```
<? php

// this code section is used to check if data for the search has been submitted.

if (isset($_GET['keywords'])) { $keywords =mysql_real_escape_string(htmlentities(trim($_GET['keywords']))); $errors = array(); if (empty($keywords)) { $errors[] = 'Please enter a search form'; } else if (strlen($keywords) < 2) { $errors[] = 'Your search term must be two or more characters'; } else if
 (search_results($keywords) === false) { $errors[] = 'Your search for '.$keywords.' returned no results'; } if (!empty($errors)) { foreach ($errors as $error) {echo $error; echo '</br>'; }}} ?>

//the following function code section is used to keywords and does the searching so that search results are then returned.

function search_results($keywords) { $keywordsRaw = $keywords; $returned_results = array(); $where = ""; $keywords = preg_split('/[\s]+/', $keywords); $total_keywords = count($keywords); foreach ($keywords as $key=>$keyword) { $where .= "`keywords` LIKE '%$keyword%'"; if ($key != $total_keywords - 1) {
            $where .= " OR "; }}

//this code section displays the search results pages, showing the title, brief description and the URL where you can click to see more information.

$query = "SELECT title, fn, descrp, url FROM articles WHERE $where"; $newResult = mysql_query($query);
            $recs = mysql_fetch_assoc($newResult); $recsFound = mysql_num_rows($newResult); $retArray = array(); if($recsFound==0){ return false; } else { echo 'Your search for <strong>['. $keywordsRaw .']</strong> returned <strong>'. $recsFound .'</strong> results:<br />'; do { echo "<br /> <strong>".$recs['title']."</strong>"; echo "<br /> ".$recs['descrp']; $splUrl = preg_split("{[\>]}", $recs['url'], 2); $splUrl2 = preg_split("{[\<]}", $splUrl[1], 2); $recs['url'] = $splUrl2[0]; echo "<br /> <a href='".$recs['url']."' target='_blank'>". $recs['url']."</a>"; $splfn = preg_split("{[\>]}", $recs['fn'], 2); $splfn2 = preg_split("{[\<]}", $splfn[1], 2); $recs['fn'] = $splfn2[0]; //echo "Show what to show in XML".$recs['fn']; echo &nbsp;| <ahref='viewXML.php?fn=".$recs['fn']."'><strong>[View XML format]</strong></a>"; echo "<br /><br />";} while ($recs = mysql_fetch_assoc($newResult));
```

Fig 2 shows lines of codes segment for searching with our EWEBSEARCH model.

For instance, let's take a practical example of a user search using the following (query, q) i.e. keywords [Web search+ semantic search, semantic Web, Web 2.0, ontology+ social annotation] to search for articles related to Semantic Web and the Web. The EWEBSEARCH model then searches, crawls and indexing, returns Response or does Retrieval and finally ranks related articles as search results based on relevance to users' needs as shown in the Search Result Page (SERP) displayed below with title of article, a brief description and URI (Unified Resource Indicator) of the article. See fig3.

```
<? xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault
="unqualified"  elementFormDefault=
"qualified" xmlns:xs
="http://www.w3.org/2001/XMLSchema">
<xs:element name="Articles">
<xs:complexType><xs:sequence>
<xs:element maxOccurs="unbounded"
name="Articles"><xs:complexType>
<xs:sequence> <xs:element name="rootEl"
type="xs:string" /> <xs:element name=
"titles" type="xs:string" />
<xs:element name ="dte" type="xs:date" />
<xs:element name="descrp" ype="xs:string"
/><xs:element name="url" type="xs:string"
/> <xs:element name="keywords"
type="xs:string" /><xs:element name=
"email" type="xs:string" />
<xs:element name="website" />
</xs:sequence></xs:complexType>
</xs:element></xs:sequence>
</xs:complexType></xs:element>
</xs:schema>
```

Fig3: Shows Schemas of XML meta-tags of some content of Web pages of EWEBSEARCH model.

7. CONCLUSION

Semantic Web is considered as Web of Data. It is not the newer version of Web but it only advocates for the conversion of existing contents of Web into machine readable form. The machines require semantics information to establish relationship among the content. The major limitation of current search engines is the lack of these missing semantics in current Web contents. This results in huge number of retrieval of results. Most of them are neither reliable nor relevant.

Following the Objective of this research work, which is to develop a conceptual model for improving Web Search using Semantic Web technology and implementing the model. We were able to present a prototype of the proposed semantic web based search engine named EWEBSEARCH model powered with XML meta-tags deployed on the web page stored in the database to search the queried information. The XML pages consist of user defined tags. The metadata information of the pages is stored and extracted from the database.

8. REFERENCES

[1] Abur M. M., Enhancing Web Search using Semantic Web Technology; M.Sc. Thesis; Ahmadu Bello University ABU, Zaria Nigeria, 2012.

[2] Alhassan Adamu (2011) thesis work: the Implementation of Semantic Web methods to Search engines.

[3] Antoniou, G. & Harmelen F. V. (2008), A Semantic Web primer 2nd edition.

[4] Berners-lee, T. (1997).”Metadata architecture.” http://www.w3.org/ Design Issues/Metadata.html,Jaunary 1997.

[5] Berners-lee, T., Hendler, J. & lassila O. (2001). The Semantic web, Scientific American, May 2001 pp. 29-37

[6] "Bing Search Engine".http://www.bing.com

[7] Evri: About Us. 2009 http://www.evri.com/about.html>.

[8] "Google Search Engine".http://www.google.com

[9] Ledford J. L;(2008) Search Engine Optimization Bible. Wiley Publishing, Inc

[10] Levene M., An introduction to Search Engines and Web Navigation. (2010), second edition.

[11] Lyndon N. and Elena P. (2004), State of the art of current Semantic Web Services initiatives

[12] Manning C. D., Raghavan P. Schütze H., (2009) An Introduction to Information Retrieval Online edition.

[13] Pollock J. T., (2009) Semantic Web for Dummies.

[14] Sensebot semantic web search engine (2010).

[15] Tümer D., Shah M. A., and Bitirim Y., An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia, 2009 4th International Conference on Internet Monitoring and Protection (ICIMP ’09).

[16] "Yahoo Search Engine".http://www.yahoo.com