A Combined Measure of Semantic Web and Meta tag technology for Web Retrieval

Sulaiman Al Rayee
Al Imam bin Saud University
Riyadh, Saudi Arabia
reyaee@gmail.com

ABSTRACT: *Increasingly web is used for tracking numerous information pieces of large and complex data. Normal way of accessing these sources using web search engines has many limitations and we have enough literature to document them. This study is an enhancement to trap the discovery of targeted information pieces which are normally failed cases in search engines. We use an integrated approach of conventional web search and semantic web by carefully constructed algorithms. The algorithms are tested to track data about the gulf institutions. The experiments provide encouraging results and enable us to draw future directions of research in this domain.*

Keywords: Web, Web construct data, Institutions Profiles, Gulf institutions

Received: 11 September 2011, Revised 7 November 2011, Accepted 21 November 2011

1. Introduction

A tremendous amount of textual and other data is being produced with the motivation of deriving insights through analysis of similarities, differences and interactions among data sets. The search of them using conventional web is generally sufficient for retrieving them that share a high level of similarity while searching a single concept. The data retrieval evaluation becomes essential in discerning more distant co-relationships. It has been proved with evidences that simple conventional search using search mechanism is not suffice.

The currently available web search methods provide database crawl with the unstructured web content, but defer a detailed structural relation among related data to further analysis by external crawl and processing methods. More importantly, they focus on finding data that share similar overall content or composition, and are not prone to sensitive to detect connected information pieces. Such structurally non-connected but semantically connected data identification is important in detecting functionally significant similar data. In this study, we propose an integrated approach to both search and crawl tasks.

It has been known that the bulk of information available on today Web is clearly disconnected, diverse and hard for common user to search in, for the desired information. Some of the popular search engines such as Google, Yahoo and Bing use a few common as well as specific algorithms to crawl the pages. Among them google is widely used than rest due to many features including the popular page rank algorithm and its extensive coverage. There are many problems in conventional searches due to the voluminous amount of information and services available. Many earlier and recent studies have documented this fact and we can postulate that the end users need of a robust aggregation and consolidated mechanism returning fewer but more reliable results [1].

In our current work, we integrated two parsing systems, the Conventional Web and Semantic Web coupled with meta tags, as they have their own strength. The conventional web provides more variety of authoritative data in more unstructured and machine un-friendly by using heuristics and machine learning algorithms for knowledge exploration. The next one, the Semantic Web has less datasets but in more structured form, which can easily be located and disambiguated. We in this study show that combining search mechanisms along with intelligent use of semantic technologies and datasets related information can be located, disambiguated and delivered to the user.

2. Related Work

The conceptual and seminal contribution the semantic web as viewed and advocated by Tim Berners-Lee et al. [2], is intended to provide a solution to the search problems. Whereas the web as we know it can be thought of as an ocean of pretty, linked, human-readable data resource, the semantic web proposes to create a global sea of rich machine-comprehensible information.

While not universal, the semantic web vision has many enthusiastic adherents; what is not clear is the concrete path that would lead to the realization of the vision. To extract meaningfully related sources, the World Wide Web Consortium (W3C) is promoting the Resource Description Framework (RDF) and the International Standards Organizations (ISO) is developing Topic Maps and other less well-known standards. We have been experiencing many efforts to secure strongly build algorithms where as a few focus on explicitly at the semantic web. The semantic web incorporates many features and tools where the data producing institutions publish data in some specialized dialect of XML. It is visible and clear that the information trapped in their documents would make a great contribution to the semantic web.

By discussing some efforts the possibilities of semantic data retrieval, we below provide the scope of our work. The significant information about an institution is available in many types and forms of web pages and related services such as in the Web 2.0 services like blogs, forums, Facebook, etc. The traditional search mechanisms such as google and others return a bulk of unsorted information; now many search options are available for finding company or insituton information. The prominent among them are Zoominfo [3], 123 people [4], Pipl [5], Intelius [6] which get popularity now. They use services adopted conventional crawling, indexing and parsing of Web pages to retrieve and cluster the information. For example Zoominfo currently crawls and indexes more than 45million people and nearly 5 million companies and institutions from the open Web to locate details of individual people and companies and then constructs profiles using artificial intelligence techniques (see [7]). Some other search machines like Intelius or Zabasearch [8] use publicly available government records and commercial sources [9] in addition.

Currently the Semantic Web with structured, interlinked, machine-readable databases is gaining more prominence. Our work differs from normal search mechanisms as listed earlier. Standard searches of users crawl web engines mentioned above (for example Zabasearch, Zoominfo, …) by using a search engine Bing Search [10]), which results we streamline using heuristics and cluster similar to people search engines, in combination with Semantic Web databases like Linked Open Data to enrich already found resources. By realizing this potential we in the current study employ the integration of Semantic web and Conventional Web (Web services, Datasets especially Web2.0) to detect information about an institution and integrate it and present it to the user. We believe that the use of this additional data which is connected and navigated with semantic resources can help us provide a more detailed, dynamic and interlinked institution profile.

3. Data Set

In the current work, we explain our approach using a data set of 187 institutions of Gulf countries. These 187 are drawn based on a combined measure of google page rank and retrieval ranks obtained from three major search engines, google, yahoo and bing. The crawling process has enabled us to search extensively and bring out the target institutions by using multilevel heuristic approach. The Gulf countries have enormous institutions specializing on different levels of educations across many domains. We have identified as many as 2849 total institutions; however the top 187 institutions are identified and categorized into 28 domains and levels. The identification of the top domains is somewhat tricky as we do not use any parameters to rank the quality or size. Our system does the retrieval of high web influence.

The Google theory goes that if Page A links to Page B, then Page A is saying that Page B is an important page. PageRank also factors in the importance of the links pointing to a page. If a page has important links pointing to it, then its links to other pages also become important. The actual text of the link is irrelevant when discussing PageRank. The toolbar is just a *representation* of actual PageRank. The following table provides the actual page rank versus the toolbar presentation

If the actual PageRank is between	The Toolbar Shows as below
0.00000001 and 5	1
6 and 25	2
25 and 125	3
126 and 625	4
626 and 3125	5
3126 and 15625	6
15626 and 78125	7
78126 and 390625	8
390626 and 1953125	9
1953126 and infinity	10

Table 1. Page rank versus tool bar

The significance of any one factor in search engine algorithms depends on the quality of the information it supplies. A factor's importance is known as its weight. To demonstrate how weighting is arrived at, it's easiest to move away from PageRank for a second and look at Meta tags. Originally, when the Meta keyword tag was new, it is possible to prepare the code as below.

< meta name = "keywords" content = "pagerank, pagerank uncovered, algorithm, algorithms" >

In theory, the Meta keyword tag was a very good indicator of what the page was about. However, as most are well aware – the weighting for the keywords tag is fast approaching nothing. [11]

Two factors have contributed to this as viewed by Ridings and Shishigin. [11]

1. The *ease* at which Webmasters can manipulate it.

2. The *level* of manipulation by Webmasters.

It is pertinent to specify that there is a high demand for gulf information form web users, the average volume of data about the institutions is high. Thus, finding information on institutions at Gulf is a difficult task. Many empirical studies can only find data which is stored in the pages of web. This can be a problem when dealing with institutions search.

4. Approach

We have developed and used the following algorithm to draw the candidate institutions in gulf.

Let *I* be the set of Web pages that can be reached by following a chain of hyperlinks starting at some root page, and let *n* be the number of pages in *I*. For Google, the set *W* actually varies with time.

Let *M* be the *n*-by-*n connectivity matrix* of the candidate pages, that is, *gij* = 1 if there is a hyperlink to page *i* from page *j* and *gij* = 0 otherwise. The matrix *M* can be huge, but it is very sparse. Its *j*th column shows the links on the *j*th page.

The number of non-zeros in *M* is the total number of hyperlinks in *I*.

Let *ri* and *cj* be the row and column sums of *M*:

$$ri = \sum_{i} gij\,;$$

$$cj = \sum_{j} gij\,;$$

The quantities rj and cj are the *in-degree* and *out-degree* of the jth page. Let r be the probability that the random walk follows a link. A typical value is $p = 0{:}85$.

Then 1- r is the probability that some arbitrary page selected and _ = (1- r)= n is the probability that a particular random page is selected.

As described above we tested our application on the Gulf institutions test set. There were 2249 institutions from different fields of research for example engineering and technology, economics, physics, management, history, computer science and other scientific fields. As search query we automatically inserted full name of the scientist from the Gulf region member database into our application.

The metric adopted here to classify these different structures in the coordinate (out-degree, in-degree) space has been applied in a more recent study [12] to discovering hubs and authorities in a variety of graph web-based pages.

We have operationalized the algorithm to generate the candidate institutions.

Table 1 below shows the example results for our first experiment with institutions search for Gulf institutions test set.

Gulf institutions test set Search	
Number of domains	
Returned intuitions per	12.22
domain without heuristic	19.068
Returned images with heuristic	13.349

Table 2. Gulf Institutions Test Set Search

During our search and optimized operations we draw search files and returns. As we can see from the results of Table 1 this simple heuristic already helps in reducing the amount of false results. It is also clear that just parsing URLs does not yet solve the disambiguation of the institutions identification. For example *Canadian University of Dubai* is an institution in UAE, but as this will likely to return large number of institutions with name university and Canada. Even the search with free phrase in google returns 2147 results for "*Canadian University of Dubai*" and the phrase search returns only 18 irrelevant results. One possible solution for this problem is adding some additional metadata to the query text like the location, address field etc in this case, however this means that user has to have some knowledge about the domain of the searched institution. In addition by restricting this application to search about various information concerning Gulf Institutions, or institutions in general, one could also profit of having a consistent domain-related metadata basis (eg. scientific metadata) and therefore also better results.

4.1 Search Engine Crawl

In our first experiment we used the search application for searching for textual pages of Gulf institutions in the Web. As indicated earlier we wrote this application in C# using Bing version 2 from Microsoft [13]. Bing Search API is used for querying the name of the searched person and it can return different types of results like Web pages, images, videos etc. The API returns various response fields for each result, see [14], amongst others it returns the Text Result. This retrieval represents the meta tag which represents the alternative text for the target URL and expected to get displayed when user navigate in the web for target or candidate pages, or when the target page cannot be displayed on in the web. In this environment, the tag normally contains a quite good representation of what is shown by the web page. In this experiment we parse the meta tag for the name of the institution in the query and we only return pages where tag contains at least one part of the name of the institution. After the query is finished, we display the top relevant pages to the user. Each thumbnail leads the user to the web page containing the institution so that the user can then have the finding whether the displayed result is correct.

The same approach can of course be used for various usages, for example also for web content. A different part of our applications does this. Given the same query (= name of the searched institution) for the content search our application starts various Web search queries using different metadata. It simultaneously performs a web search for all information such as departments, courses, curriculum, publications, research, etc. A search for other content items in the Web is also performed. In this case we parse the returned results so that and return only results where the URL contains at least one part of the name given in the query in order to get high quality results.

4.2 Semantic and Ontology Crawl

In our Semantic Web and ontology approach we used customized search profile called as CSP as our focused source to find web pages of the target and candidate institutions. Our basic dataset is used to specifically have the publication list to add more variety in the anticipated profile of the institution. In semantic web data a specific and unique URL is used to represent an institution, so locating a URL is considered as an important task. We have developed the appropriate algorithm which uses semantic as well as ontology approach and tools to crawl the URL. After applying this algorithm on the web, we have identified 2249 members. In the end we able to find 1798 CSP of URLs and 155 CSP from target institutions and 505 DBLP URI of the members as listed in Table 3.

Academia Europea test set Linked Data Search	
Number of domains	16.36
CSP of URLs	1798
CSP from target institutions	155

Table 2. Result returned in Semantic web search

In a summarized discussion of our two approaches, for future we envision organizing this information and presenting the results as shown in Figure 1 below: here we have grouped targeted text by Bing search and semantic data retrieved information in one profile. We hope that this proposed system will give user a coherent and detail view of the information at one place. Still this initial aggregation returns a large set of links and documents and we believe that some sort of editorial process is needed due to sort out and consolidate this information!

The editorial process is also due to copyright issues it is important to think about how the consolidated information will be saved and presented to the user. It is not just possible to gather all the information from different sites and copy it to a consolidation server. This means we need to create a possibility to consolidate and display relevant data without infringing copyright. Instead of the trivial approach which would be to verify whether found pictures, videos and other multimedia content may be used, here we propose a tool where an editor can easily prepare steps necessary to gather and consolidate the information wanted and display it on-the-fly when user requests the information.

CANADIAN
UNIVERSITY OF DUBAI
Design Future Environments!
or holding a diploma in design, construction and related discipline?
credits towards our new 5 year Bachelor program in Architecture or in Interior Design
Register now!

Figure 1. Target retrieval top rank in the combined search

Our suggested approach is to gather a first version of consolidated information for a institution with all relevant and unique links offered by the combination of Metatag and Semantic Web. In the second step we ask a team of editors to go through the links, check them and annotate them. For example this editorial team first looks at the links and thinks about how to display relevant links and data on the server without infringing copyright. Then an editor uses a special plugin and stores in it the steps how the data of the interest is to be handled and displayed. For example some pages are copyrighted and hence drawing all pages is somewhat difficult. Thus when we consider a candidate institution as the target, the algorithmic part of our approach returns a link to the text of pages of an institution on, among other pages search. In order to be able to use this analogy the searchers or developers use a plugin. This kind of plugin takes a group of parameters describing what to do in this case, for example do a screen dump, scroll to a part of page and cut out the picture using given coordinates, the date of link retrieve etc. First author described a similar approach in [15] by developing a presentation tool called SIP which is able to avoid the violation of copyrights, allow access to data that may not be available to all and assure that the sources are always quoted properly.

We believe in using semi-automated approach for information retrieval in our application where firstly given query results from web content from a combined institutions profile for relevance of the results and copyright issues.

5. Conclusions and Future Work

We discussed an integrated approach for combining results from the conventional web search and Semantic Web into a structured institution profile. In the future this approach could also be used for aggregation of results for many other objects and concepts.

While we apply semantic web using the algorithms we have decided to introduce more scalable and other related applications which can handle multi-platforms in the complex web.

We would like to specify that our approach can be further refined as we address the search and retrieval in one angle. Newer research methods have been created in semantic web and it is important to consider and apply in an integrated manner.

References

[1] Alexander Korth, A. (2009).The Web of Data: Creating Machine-Accessible Information, Read Write Web, 18th Aprir http://www.readwriteweb.com/archives/web_of_data_machine_accessible_information.php (Last visited: 3.09.2010).

[2] Berners-Lee, T., Hendler J., Lassila O. (2001). The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" Scientific American Online Edition, 17th May , http://www.ryerson.ca/~dgrimsha courses/cps720_02/resources/Scientific%20American%20The%20Semantic%20Web.htm (Last visited: May 2011).

[3] Zoominfo http://www.zoominfo.com/ (Last visited: 3.09.2010).

[4] 123people http://www.123people.at/ (Last visited: 3.09.2010).

[5] Pipl http://pipl.com/ (Last visited: 3.09.2010).

[6] Intelius http://www.intellus.com (3.09.2010).

[7] http://www.zoominfo.com/About/company/technology.aspx (3.09.2010)

[8] http://www.zabasearch.com (Last visited: 3.09.2010)

[9] Ramasastry, A., Can We Stop Zabasearch – and Similar Personal Information Search Engines?: When Data Democratization Verges on Privacy Invasion, http://writ.news.findlaw.com/ramasastry/20050512.html (visited: 3.09.2010)

[10] Bing API, Version 2, http://msdn.microsoft.com/en-us/library/dd251056.aspx (Last visited: 3.09.2010

[11] (http://www.voelspriet2.nl/PageRank.pdf) (PageRank Uncovered Formerly PageRank Explained Chris Ridings and Mike Shishigin. VERSION 3.0 Last amended – September (2002).

[12] Google's Web Page Ranking Applied to Different Topological Web Graph Structures, George Meghabghab. *Journal of the American Society for Information Science and Technology*, 52 (9) 736–747, 2001

[13] Dbpedia.org (Last visited: 16.02.2011)

[14] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2010). Harvesting Pertinent Resources from Linked Data, *Journal of Digital Information Management*. 8 (3) June, 205-212.

[15] Trattner, C., Helic, D., Korica-Pehserl, P., Maurer, H.(2010). Click, Click - sand an Educational Presentation is Available on the Web, *In:* ED-MEDIA.