Enhance the content search by using Semantic Web

Aeman Jamali
MS Computing,
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600

Asim Riaz
Assistant Professor, Department of Computing
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600

Abstract— **-- The content search of web content for large data and information presents enormous resourcing and quality challenges. Users expect to find information quickly, with minimal navigation and with consistency of information and nomenclature. For example content and solutions information, users expect clear, relevant lists of information and services that comprise those solutions, including research papers, publications, videos, images, interviews, conferences and case studies that provide referential examples. The foundation component of the Semantic Web is ontologies which are used to embody knowledge in the Semantic Web. Ontology is an information form is used to demonstrate a set of concepts and the relationships connecting those concepts within a domain. Taking unstructured data from the web and formalizing it so that it can be structured automatically is a difficult work to do, but apart from its significance it is interesting as well. Through this research it is intended to make the automation of ontology public, as it is based on open standard and constructed using publicly available resources of Google, like Google AJAX API and JavaScript parser JSON.**

Keywords: Ontology, Resource Description Framework (RDF), Metadata ,Web Ontology Language (OWL) ,World Wide Web (WWW), internal structure (IS), Information Retrieval (IR).

I INTRODUCTION

Today, people of the world are dependent on World Wide Web search engine, and changes on these search engines can be seen weekly. Those of us who have to get real work done using these engines just want to know which ones we should use when, and what we should know about how they work [1].

People of the world are dependent on worldwide web. A proper search engine consists of database and tools which generate database .Most search engine allow you to type your query and find out appropriate result for you. But some of us are not satisfy by that results, because sometimes that result is not related to query we asked. Semantic web consist of Syntax means "meaning behind the sentence". But still semantic web content search is not enhance .In our article we have introduce Enhance Semantic search ESS technique to enhance the content of Semantic web which improve quality and consistency of the web.

A. Ontologies in Semantic Web

Ontology is basically representation of the knowledge segregated in sets, so the concepts of same domain can be easily understood [10], and it is the vocabulary extension of Resource Development Framework (RDF). Knowledge and concepts are inter-related with each other, for acquiring the knowledge we need to clarify the concepts, and for clarification we need to understand the actual context of data. Ontology is discovered for the same purpose, so that web search can become easier for everyone. Ontology is a Keywords representation is used to demonstrate a set of domain keyword concepts and the association within that domain keyword [10].

B. Searching on the Semantic Web

There are several Semantic Web search engines e.g. Kosmix, Swoogle, Factbites, Exalead, Power set, Sensebot etc., they are used to explore and retrieve the ontologies from the Web but most of the search engines are not yet put into practiced since they are in research level. The Semantic Web search engine can be roughly divided into two categories; one is Ontology search engine and other is Semantic search engines. Our effort is to find out Ontology search engines which are exercised to find the Semantic Web documents; they use a technique named ranking technique to find out the closest results of the query which user asked for search [3].

C. Semantic Relatedness

In this paper we introduce a solution to the very common problem with Ranking pages on the web and finding out the weight of the page to estimate the relevance of documents with respect to query which is being asked by user. Enhance Semantic search (ESS) ability are needed to overcome the limitations of long-established search engines that are mainly keyword based search engines. Ontologies play main role to attain this goal. Ontology is defined as "ontology defines as an explicit and formal specification of shared conceptualization". In our work ontology can be used as set of terms and relationship used between different concepts [1].

II. BACKGROUND AND RELATED WORK

There is a lack of automatic and well-grounded methodologies on web to calculate and weigh ontologies [4]. Ontologies may be reviewed and analyzed from different angles, such as how the

ontologies can be rated and reviewed by users , how well they meet up the requirements of certain evaluation tests and results [5]. Gangemi and his some colleagues [6] defined three main types of evaluation to measure ontologies; which are functional, usability-based, and structural evaluation. Among these functional evaluation considers the measuring of how well an ontology is serving its purpose and use [6]. A usability evaluation is related with metadata and annotations [7]. Structural evaluation focuses on the structural properties of the ontology the same as a graph.

Harith.A and Christopher.B described the purpose of the testing and evaluation is to make a system to rank ontologies returned through search engines according to how well the ontologies carry out under certain measures. Google commonly uses the PageRank method for ranking of documents presented on the web. Some ontology search engines implemented a PageRank-like method to rank ontologies by studies links and referrals between the different ontologies in the expectation of identifying the most popular ones (Swoogle [8, 9] and OntoKhoj). On the other hand, the mainstream of ontologies available on the Web is unsuccessfully connected, and more than half of those ontologies are not referred to by any other ontologies at all [5]. Bad connectivity would certainly produce bad PageRank results

.

In addition, a popular ontology does not essentially point out a good quality representation of all the concepts it covers. For illustration, supposing an engineer was searching for an ontology about "students", so there might be an ontology about the academic domain that is close connected to the ontology named students, and as a consequence popular. If this ontology holds a concept model named "Student", then this ontology will let that engineer see up high on the list of candidates. Though, it may very well be the scenario that the "Student" class is very imperceptibly represented. That ontology could turn out to be popular due to its coverage of publications, papers and research topics, rather than for its coverage of student related concepts. One more challenge for ranking the ontologies is looking for different multiple terms. For instance, if looking for "pet" AND "food", afterward an ontology that has such classes in good structural proximity to each other is enhanced than one where those classes be further apart. Different formulae exist for ranking in the text to measure similarities of terms within the semantic networks, and these be able to be used uniformly to measure structural proximity.

This paper conducted test with a adapted set of ranking measures to those we before used and described in.

Antonio M. Rinaldi, he describes Semantic relatedness which is used to evaluate the relevance of documents on the web with respect to query which is asked by user. In his DYse approach he considered three major techniques, which are metrics, semantic relatedness and WordNet. Metrics are applied on Semantic relatedness to find out the relationship between two words. WordNet is used for dictionary as well as to measure the semantic relatedness of two nouns related to each other and also find distance between them by keeping subject keyword and domain word. This approach is applied on Dynamic semantic network (DSN).

III. METHODOLOGY

A. Architecture

A The methodology used in exploring URLs for constructing the final ontology along with the selection of the class names is describe here

..

The methodology used in exploring URLs for constructing the final ontology along with the selection of the class names is describe here

.

Figure 1 demonstrates the detailed mechanism for analysis of websites in huge amount, so that the domain can find its related concepts by going through the inter-related keywords. To find the output with most accuracy JSON is used for its processing. Further to it, final ontology is constructed using 2 elements: first is, classes: selected concepts, and second is, OWL: language. Each concept is associated with the URL where the concept is extracted from. To present the most feasible ontology hierarchical order, the said process is run recursively to achieve the appropriate results [14]

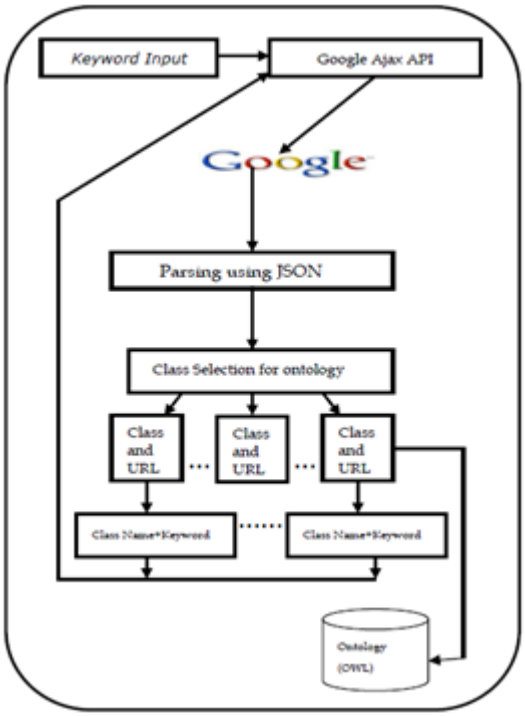

Fig 1: Main Architecture

The following sections explain the process in detail.

B. Google Ajax API

Figure 2 depicts the architecture for interaction between Google Web Services and front-end application. Application developers have choice to use any programming or scripting language (Java, .Net, php, Python, Perl) they are comfortable with to build the connection with Google Web APIs services available remotely. Users' queries for searching/extracting any information are

processed in Google Server. And all of the above communication is done through Google AJAX Search API, as there is very less coding involved in integration on web page of Google' search mechanism and its controls.
These include [10]:

1) Web Search: This is common as everyone uses it for searching the desired information from web, they simply enter their queries, and they get a list of search results on web page.
2) Local Search: The searching here is performed on specific location using Google Map.

3) Video Search: Video search results are extracted using the AJAX Video Search. As soon as it is builds its connection, the application will process search requests, and those requests are after processing in Google's index and then spell check is performed in Google cache, the structured and accessible information is produced.

The basic functionality of the AJAX APIs is the integration of the hosted services with customized web pages, it allows this through JavaScript code, and for this reason Google's widely known hosted services like Google Search and Google Maps are enhanced, and can be directly accessed by anyone.

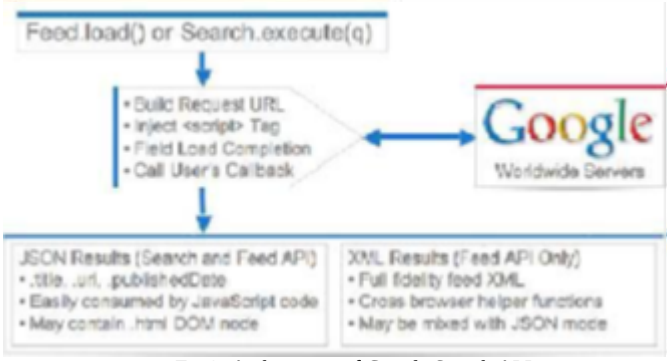

Fig 2: Architecture of Google Search API

The core JavaScript code's methods for searching is Search.execute ( ) and for feeding is Feed. Load ( ). Once the request is received at Google server, the above mentioned methods are executed and response is generated on web page either using JSON or XML formats. On the other hand, the parsing is done in either way manually or automatically through provided UI controls of AJAX APIs.

C. JSON

JavaScript Object Notation (JSON) as shown in Figure 3 is designed for making data easily readable for humans without creating any heavy process as it is text-based open standard. Its derivation is as its name indicates from the JavaScript programming language to make data structures and objects simpler. Though it is associated with JavaScript, but still it has parsers available thus making it language independent (i.e. any programming language can utilize it). Figure also shows on of these forms that the String data structure can take. JSON Schema defines the structure of JSON data, and how it can be utilized in particular application and how it can get modified accordingly, basically it is specification for JSON-based format. Its concept is taken from XML Schema which is used for XML format, and provides features such as documenting (self-descriptive), validating, and interacting JSON data [10].

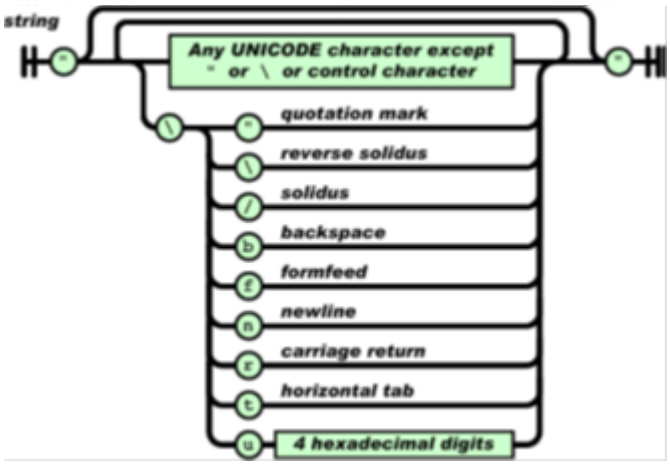

Fig 3: JSON Schema

The system we are proposing through this research has used JSON to great extent in parsing the Google API response, as it gets less complex, and through JSON system can store the results in array, then process with retrieving the desired URLs, its content, and count result. It has also helped in the analysis of the candidate words [10].

D. Ontolotgy Representation

Ontology is basically representation of the knowledge segregated in sets, so the concepts of same domain can be easily understood [10], and it is the vocabulary extension of Resource Development Framework (RDF). Knowledge and concepts are inter-related with each other, for acquiring the knowledge we need to clarify the concepts, and for clarification we need to understand the actual context of data. Ontology is discovered for the same purpose, so that web search can become easier for everyone. Ontology is a Keywords representation is used to demonstrate a set of domain keyword concepts and the association within that domain keyword [10].

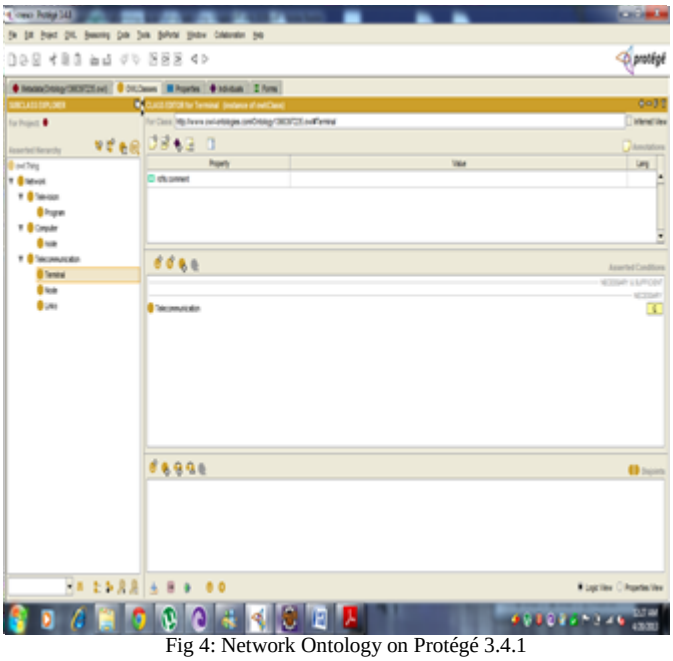

Fig 4: Network Ontology on Protégé 3.4.1

Protégé has plug-in called Jambalaya [10] that provides user with graphical presentation of the visualized hierarchy (Jambalaya extension can be easily installed with Protégé). Jambalaya uses Shrimp for visualization of Protégé -Frames and Protégé –OWL ontologies. Refer Figure 4 for ontology with URLs for the keyword Network.

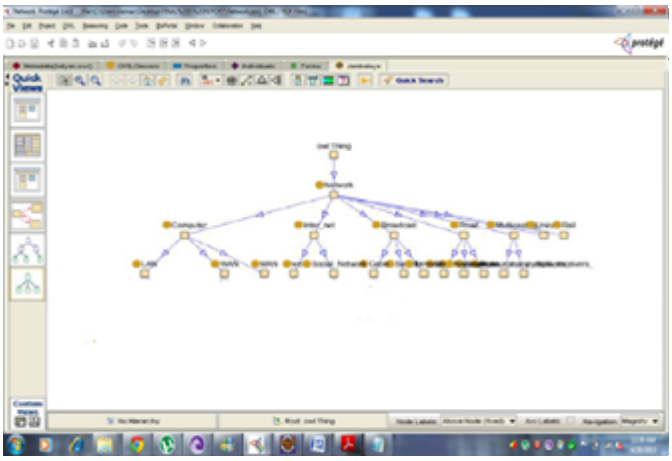

Fig 5: Jambalaya tab for Network Ontology on Protégé 3.4.1

IV. Design and Implementation

Following procedure has been followed for the implementation of the application using Java and JSP. The detailed architecture is been discussed in context of Figure 2.

1) The process begins with choosing a keyword Network, for instance and then entering it keyword in the HTML interface, for which the ontology is basically constructed;

2) Once we provide the input in terms of keyword, the program uses Google AJAX Search API for output i.e. to retrieve the information of that keyword along with the URLs that contains that keyword.

3) The result for our entered keyword Network is shown in Figure 6. The provided result (data) is in the form of an array that includes all useful information of the matching websites, titles, URLs, etc. The response date is mentioned below, filtering is applied over it, for example, putting restriction on the number of returned results (100 is the number in our case), and many other filters are used in the program.

Fig 6: Network keyword result after searching in Google Ajax API in array output

4) For the construction of the ontology, data is required to parsed, so that class and URL selection becomes more appropriate. JSON has been used for parsing the response data. Parsing makes it easier for the program to split the websites retrieved against the keyword provided and also capture relevant results for the keyword (Network).

5) URLs selection for the class is dependent upon the occurrence of the keyword in the content that how many times it has been used in that particular webpage. We have extracted this content from the result data with using JSON as mentioned earlier. Appropriate URLs are selected and then classes and sub-classes are defined for the representation of the hierarchy of the desired keyword and its association accordingly. Then it is programs function to determine the relevant words with the main keyword, which means that it does not contain prepositions, etc. and their size does not exceed two characters as well as they are represented in standard ASCII.

6) Each resultant word (candidate) selected is passed through an analysis that includes checking the number count of its occurrence in the web content. After performing analysis,

total number count is noted and on that basis an appropriate candidate key is selected.

7) When we have got the resultant word or candidate word, a new keyword is formed by joining two words; candidate word and the main keyword, Network Computer Network, for instance. Similarly the whole process can be repeated, each recursion can have its own selected candidates, but keeping in view the above mentioned constraint. The recursive process continues till only when there are no results left to be found for the word.

8) We have got the final output in graphical representation of hierarchy of class and sub-class and that is our ontology.

V. EVALUATION AND RESULT

The initial keyword "Network" is selected here for demonstration, and it has been carried along with the constraints mentioned earlier, that minimum size of the chosen word shall be more than two characters, and occurrences (number count) on the web content are two of the few constraints. It can be seen in figure 8, which is visualization of the class hierarchy is protégé. To further elaborate the example, another word i.e. the candidate word computer is taken, so the combination of initial and candidate word is "computer network", and accordingly the result is produced from various web sources. To analyze it further, the hierarchy is built with different candidate words from broadcast (mainly its types), such as terrestrial, cable, and satellite. The purpose of this research is building the hierarchy of classes using OWL, as it enables to find equalities or inclusions [10], so what we get in result is Internet class contains social network and similarly social network will present the output that is node. Following figure 8 depicts the visualization of OWL file in its editor Protégé.

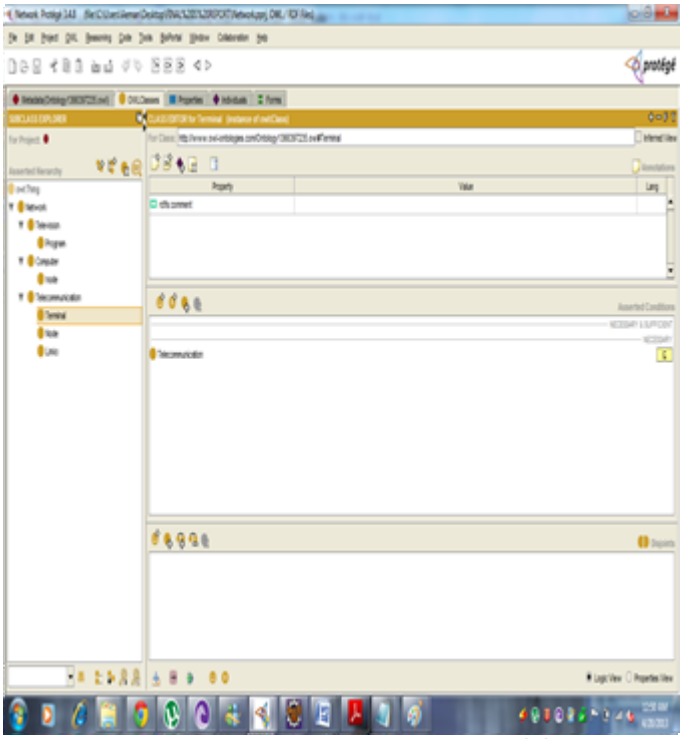

Fig 7: Network hierarchy as shown in Protégé

This system also stores the relevant URLs and the class names as well, so that user has the comfort for accessing the more relevant websites for the desired keyword. To make the process clearer figure 7 shows the store URLs along with the classes, like mail is subclass of web as well as social network. Jambalaya plug-in of OWL editor Protégé illustrates the complete ontology is various formats. Figure 8, 9 shows the nested tree map, and class tree map.

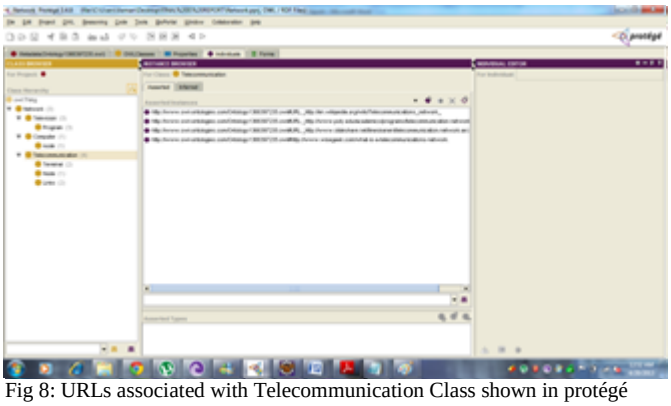

Fig 8: URLs associated with Telecommunication Class shown in protégé

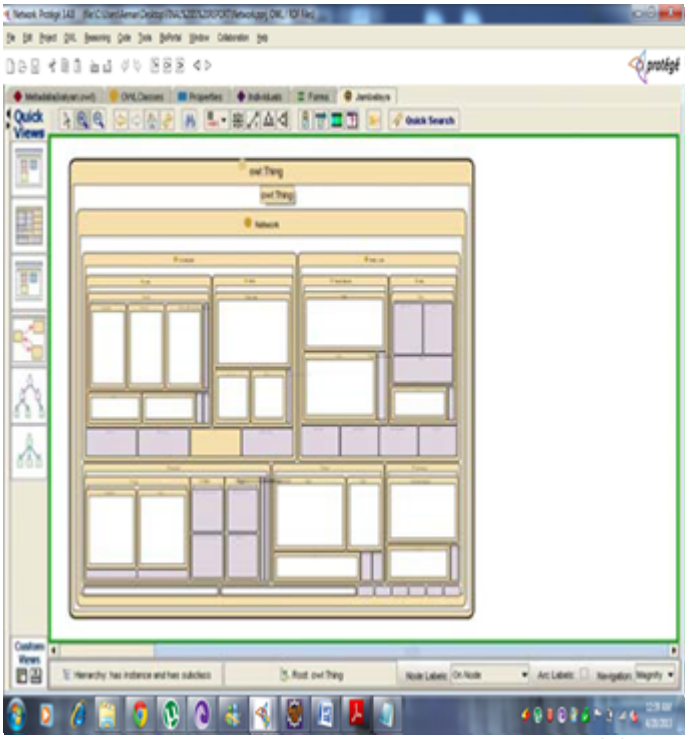

Fig 9: "Nested Tree Map" as shown under Jambalaya plug-in in protégé

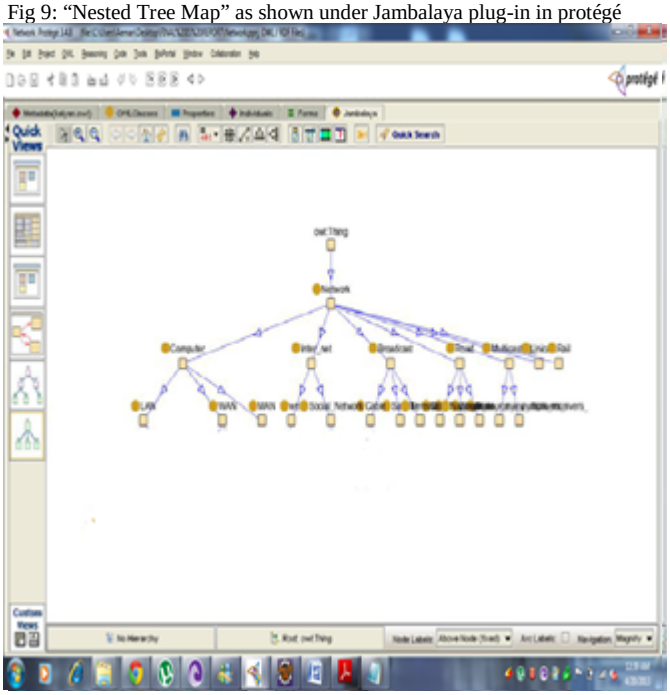

Fig 10: "Class and Individual Tree map" as shown under Jambalaya in protégé

VI CONCLUSION

Ontology construction is a growing trend and need of the semantic web as well, and many researchers is working over it in different domains, to name a few structure information working domains, databases, dictionaries, etc., and simultaneously others are getting involved in natural language texts (NLT) processes. Databases play a vital role in construction of ontologies, because they are the primary structured information sources. Taking unstructured data from the web and formalizing it so that it can be structured automatically is a difficult work to do, but apart from its significance it is interesting as well. Through this research it is intended to make the automation public, as it is based on open standard and constructed using publicly available resources of Google, like Google AJAX API and JavaScript parser JSON. The analysis done is here requires more efforts and it can be enhanced through designing algorithms and building complicated relationship of initial and candidate words.

Acknowledgement

I am heartily thankful to my supervisor, Asim Riaz, whose encouragement, guidance and support from the initial to the final level enabled us to develop an understanding of the research report. Lastly, I offer my regards and blessings to all of those who supported us in any respect during the completion of the research article.

References

[1] Bruce Grossan, "Search Engines", WebRef, February 21, 1997.Retrievedfrom http://www.webreference.com/content/search/

[2] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna, F. "Semantic Annotation for Knowledge Management", 2006, Journal of Web Semantics.

[3] Kahan, J., Koivunen, M.R., Prud'Hommeaux, E. and Swick R.R. Annotea, "an Open RDF Infrastructure for Shared Web Annotations", 2001, 10th International World Wide Web Conference.

[4] K. Supekar, "A peer-review approach for ontology evaluation", July 2005, 8th International Protege Conference pages 77–79, Madrid, Spain.

[5] N. Guarino and C. Welty. "Evaluating ontological decisions with ontoclean". 2002, Communications of the ACM.

[6] Reeve, L. and Han, H. "Survey of Semantic Annotation Platforms". Proceedings of the 2005 ACM Symposium on Applied Computing.

[7] A. Lozano-Tello and A. Gomez-Perez, "A method to choose the appropriate ontology", 2005, Journal of Database Management.

[8] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna, F. "Semantic Annotation for Knowledge Management", 2006, Journal of Web Semantics.

[9] Kahan, J., Koivunen, M.R., Prud'Hommeaux, E. and Swick R.R. Annotea, "an Open RDF Infrastructure for Shared Web Annotations", 2001, 10th International World Wide Web Conference.

[10] Netti1, "Interactive Guided Online/Off-line search using Google API and JSON", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010