International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 4, Sep 2013, 139-150

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

TOPIC SPECIFIC CONCEPT MATCHING BASED WEB SEMANTIC SEARCH ENGINE

SHRUTI KOHLI[1] & SONAM ARORA[2]

[1]Professor, Department of Computer Science, BIT, Noida, Uttar Pradesh, India

[2]Assistant Professor, Department of Computer Science, ABES EC, Ghaziabad, Uttar Pradesh, India

ABSTRACT

Keyword based search is useful especially to a user who knows what keywords are used to index the images or documents and therefore can easily formulate queries. This approach is problematic, however, when the user does not have a clear goal in mind, does not know what there is in the database, and what kind of semantic concepts are involved in the domain. The Semantic Web is an extension of the current Web that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers. Ontology is one of the most important concepts used in the semantic web infrastructure. Concerning with the users online all over the world, before planning any trip they look for various hotels available in their destination location and the facilities they provide.

But when the user sits online and searches for hotel images, they find it hard to select from images that which one is relevant and which one is not relevant. As an initial step, this paper implements a tool which makes use of Semantic Web for searching the images of hotels and displaying the results in ranked order based on user behaviour. Semantic web refines the search in such a way that only relevant images are returned.

The soul idea is to define ONTOLOGIES for various hotels along with their locations and the facilities they provide. To gather information from various hotel websites we can use EXTRACTURL tool. Ontologies can be build from this gathered information using PROTEGE tool. User enters his query from an interface and corresponding SPARQL query is generated. This query is searched in the Ontology using JENA API.

KEYWORDS: OWL, RDF, Search Entry, Search Refinement, SPARQL

INTRODUCTION

The Semantic Web is an extension of the current Web that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers. On the Semantic Web, information is described using a new W3C standard called the Resource Description Framework (RDF). Semantic Web Search is a search engine for the Semantic Web. Current Web sites can be used by both people and computers to precisely locate and gather information published on the Semantic Web. Ontology is one of the most important concepts used in the semantic web infrastructure, and RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) are two W3C recommended data representation models which are used to represent ontologies.

The Semantic Web will support more efficient discovery, automation, integration and reuse of data and provide support for interoperability problem which cannot be resolved with current web technologies. Currently research on semantic web search engines are in the beginning stage, as the traditional search engines such as Google, Yahoo, and Bing (MSN) and so forth still dominate the present markets of search engines. First, they do not provide the factor of reliability as the user demands. For example, when a particular user issue any query like “give me the images of all hotels in Delhi” the search engine although provides thousand of result to the user but it’s difficult for the user to find out which source is

reliable. The user has to sift through all the retrieved pages to find only the reliable results. Secondly, the relevancy of provided results is not up to the mark.

The web today enables people to access documents (HTML pages) and services on the Internet. Semantic Web is web of documents/belongings/concepts, things and concepts can be anything in the world – a movie, a picture, a person, a disease etc. The W3C Semantic web activity states that "The Semantic Web is a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purpose, but for automation, integration and reuse of data across various applications".

The main intent of the Semantic Web is to give machines much better access to information resources so they can be information intermediaries in support of humans. Information retrieval in current services relies only on keyword searches using Google or based on simple metadata such as that of an RSS. Moreover, there is no function to generate personalized searches easily, so users need to consider and enter search keywords that suit their own interests appropriately.

Such a keyword search is time consuming and troublesome. Moreover, users cannot perform a keyword search if they do not understand what they want to search for to some degree beforehand. Thus, when keywords cannot be specified, information retrieval cannot be performed even if users might become interested in a topic. To overcome some of these problems an alternative approach in web searching is given in this paper, wherein the user need not think of appropriate keyword that might give them the result they want, instead the user can simply provide the search engine with whatever information it has by selecting topics. The way search engine accepts user query is very user friendly and easy to understand both by the user as well as by the machine.

As concept based search is a search interface paradigm based on a long running library tradition of faceted classification. And efficient search systems have proved the paradigm both powerful and intuitive for end – users, particularly in drafting complex queries. Thus, topic – based search presents a promising direction for semantic search interface design, if it can be successfully combined with Semantic Web Technologies.

Concept based search engines differs from traditional search engines such as Google, Yahoo! Or MSN, only in the information it aggregates, index and use to answer users queries. Instead of using human readable documents such as HTML, PDF or DOC, Concept based search engines will use semantic web documents (RDF, XML, OWL).

The characteristic of these documents is that they describe things. In fact, these documents can describe anything: a person (its interests, its relation with other people etc.,), object like books, music CDs, Projects (computer projects, architectural project, etc.), geographical locations such as countries, cities, mountains, etc. So, with semantic web documents, one can describe virtually anything.

The core idea of concept based search engine is to describe query in the form of topic description. For example, a user query is "give me the images of all hotels in Delhi". These queries are not built using natural langue (such as phrases), but with an easy to use user interface that help users to build the queries they want.

PROBLEM DEFINITION OR FORMULATION

Keyword based search is useful especially to a user who knows what keywords are used to index the images and therefore can easily formulate queries. This approach is problematic, however, when the user does not have a clear goal in mind, does not know what there is in the database, and what kind of semantic concepts are involved in the domain. Experts have developed various technologies for refined search but unfortunately till now irrelevant links are got.

Consider Figure 1

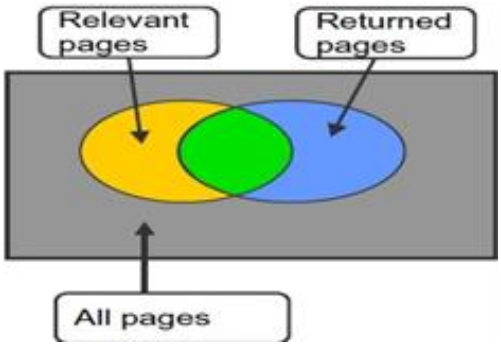

Figure 1: Precision vs. Recall

- **Precision:** Percentage of returned pages that is relevant. (green/blue) or in other words the capability of minimizing the number of irrelevant links returned to the users.
- **Recall:** Percentage of relevant pages that is returned. (green/yellow) or in other words the capability of maximizing the number of relevant links returned to the users.

All search mechanism till date performs the function where precision and recall percentage is too low.

Concept matching using Semantic Web aims to provide better precise and recall rates as compared to keyword based search. User need not think of appropriate keyword that might give them the result they want, instead the user can simply provide the search engine with whatever information it has by selecting options.

As an illustration, consider that User enters query "give me the images of all hotels in Delhi". On Google, the search results contain lots of unwanted image result which are of no interest to the user.

Figure 2: Google Search Results for Hotels in Delhi

The challenge is to create a topic based semantic web search engine which is highly user friendly and provide advance search options with the help of topics. A user shouldn't be aware of the concepts supporting the semantic web to use it. Their experience should be as close as the one they currently have with the current web and the search engines they use daily.

Some of the latest works relating semantic areas are:-

WANG Yong-gui, JIA Zhen [1] gave introduction on Web Mining and Semantic Web-related knowledge and then integrated both of them to improve the effectiveness of Web Mining. They claim that knowledge of Semantic Web makes Web Mining easier. They gave a 5 step process under the framework of Agent for integrating them.

The First Step: In the beginning, you need to build an initial ontology. To build an initial ontology first need to obtain the relevant set of atomic concepts, we use clustering algorithm to obtain the document from the Web; and then get this concept hierarchy by a variety of different ways.

One way is to use the knowledge acquisition methods to generate, such as ONTEX (ontology exploration) which input a group of concept sets depending on knowledge acquisition techniques of properties detect, and then output the level of above concept collection. Another way also can use many of the ontology models that the current ontology researchers have developed.

The Second Step: Resource acquisition module collects task-related data sets according to received tasks instructions by ontology Agent from a Web mining. Usually this step is essential. Because the data set on Web is very scattered, dynamic and often inconsistent data, whether the data collection is good or bad will have a direct impact on the results of Web mining.

The Third Step: RDF clustering module achieves ontology clustering learning to the data that resource acquisition modules has collected. The resource nodes of closest characteristics will be got together in the RDF data repository.

The Fourth Step: Data stored in the RDF data repository are mined by Semantic Web Mining module and the mining results are provided to ontology Agent.

The Fifth Step: Ontology Agent completes semantic filtering and clustering of processing for results obtained by Semantic Web Mining module, to improve the relevance of return information; and also ontology learning can take advantage of the semantic Web mining modules to carry out the expansion and modification of ontology knowledge.

Jiang Huiping [2] proposed a semantic web search model to enhance efficiency and accuracy of IR for unstructured and semi structured documents. He used Ranking Evaluator to measure the similarity between documents with semantic information for rapid and correct information retrieval. He introduced Search Arbiter to judge whether the query is answered by Keyword based Search Engine or Ontology Search Engine. He gave just a conceptual architecture of Semantic Web Information Retrieval System.

Search Engine Modules

Cache

The goal of cache is how to reduce net accesses to data sources as much as possible. When the query can be answered from cache, then the result will be directed to users without further processing by search Arbiter.

Syntax Checker and Parser

Query Arbiter and this module are designed to run in parallel contemporary. In this way, the system's processing ability will be more efficient. At first, this module standardizes the RDL queries issued by users, and then analyses whether their syntax are accordance with RDL syntax or not. If so, this component creates executable plans by a series of rules and sends these plans to Query Arbiter. Here, this module first looks up concerned information from Cache, if that information cannot be found in these buffers, it has to retrieve information from database.

Search Arbiter

In their system, they share with the idea of Mayfield and Finin[16] that semantic should be a complement of keyword based search as long as not enough ontology and metadata are available, which is done by Search Arbiter. That is,

when there is not enough ontology to answer the user's query, then the Search Arbiter forwards the query to Keyword-based search engine. Otherwise; it forwards the query to Ontology Search Engine. Note that such process need to the help of RDF knowledge bases.

Ontology-Search Engine

The component of this module returns a set of tuples that satisfy the query processed by Syntax Checker and Parser. If the tuples are only made up of instance of domain concepts, the engine follows all outgoing annotation links form the instances, and collects all the documents in the knowledge based that are annotated with the instances.

Keyword-Search Engine

When this is not enough ontology to answer the query, then the module will answer the query by tradition keyword based search method.

Extracting Agent

This module is responsible for extracting related web page from the WWW.

Wrapper

This module is responsible for returning XML document based on the material which is extracted by Extracting Agent.

Automatic Classification

This module is responsible for generating RDF documents and then it stores such documents into RDF knowledge. In additional, it will rank the documents before storing the document.

Ranking Evaluator

Once the list of document is formed, the search engine computes a semantic similarity value between the query and each document, which is done by Ranking Evaluator. With the paged ranking, the answer will be more precise.

Saman Kamran, Fabio Crestani [3] proposed a method for building a reliable ontology around specific concepts by using the immense potential of active volunteering collaboration of detected knowledgeable users on social media. They defined a seven step model for creating the semantic relationships between related concepts by using user's input from reliable well known social networking sites. They used automatic information retrieval methods called Wikipedia Link Based Measure (WLM) and Cosine Similarity for computing semantic relatedness score. This model is under construction and there is not any evaluation available yet.

In the first step we start with crawling Wikipedia pages by defining a page as an initial page that is related to a specific concept. Selecting the initial Wikipedia page for crawling, which is also the initial concept or node for constructing our semantic network, is based on prestige and authority of its page on social media.

In the second step we are building the link list of the initial page (called list A) and the link lists of pages with in/out links from/to the initial page (called list B) can be a sample of list A or B). The third step will be building lists (called Lists C) of common in/out links between List A and every sample of List B generated from the second step. We also define the ratio of the number of common links in every sample from List C and the total number of links, which exist in both Lists A, B that List C is made from them.

The fourth step will be scoring semantic relatedness between linked pages by considering common link lists. To do this, we use a combination of scores that we get from (i) WLM model and (ii) Cosine similarity model . Finally, we rank pages based on the measured similarity.

The fifth development step is suggesting concepts, whose page has a high semantic relatedness rank, to the users to whom concepts are related on a platform. This platform enables users to define a new relation and vote for the defined semantic relations by the other users among automatic suggested concepts by the system.

The sixth step is dynamically updating scores of inferred relations as well as existing relations defined based on users votes.

Finally in the seventh step we are adding discovered semantic relations, and updating scores of the existing ones by modifying: 1. the OWL ontology representation, and 2. the graphical ontology representation

Eero Hyv¨onen, Avril Styrman and Samppa Saarela [4] in their paper considered the situation when a user is faced with an image repository whose content is complicated and semantically unknown to some extent. Here ontologies can be of help to the user in formulating the information need, the query and the answers. This paper has used the ontology of promotion event to annotate the promotion event images. They have proposed the in depth structure of Promotion event ontology and the interface developed for handling users query.

The promotion ontology describes the promotional events of the University of Helsinki and its predecessors, the Empirial Alexander's University and the Royal Academy of Turku. The top-level ontological categories are also suggested. The classes of the ontology represent people of different roles (Persons, roles, and groups), events and happenings (Happenings) that take place in different locations (Places), physical objects (Physical Objects), speeches, dances, and other performances (Performances, Performers, Creators, and Works), and a list of all promotions from 17th century until 2001 (Promotions).

The main goal of the ontologization process was to create ontology suitable for the photograph exhibition and to offer the programmers the basis to implement the exhibition, either on the web or as an internal information kiosk application.

The stable and unchanging things of the subject domain, i.e., continuants, are presented with classes in the ontology. The changing things, occurrents, are presented with instances. For example, the Cathedral of Helsinki has its own class. On the other hand, buildings that are not regularly used in promotions do not have a subclass of their own, but are instances of the general class Buildings.

The instances of the ontology have literal-valued properties, such as name of the person. These properties are typically used to provide a human-readable presentation of the instance to the user. Each instance, e.g., a particular person, is related to the set of promotions in which the instance occurs. In this way, for example, the persons performing in a particular promotion are easily found.

Ontology Construction

There are several partly conflicting goals to keep in mind when designing the ontology. The ontology not only should be semantically motivated, but also easy to construct and maintain to the ontologist. At the same time, the annotation work based on it should be simple to the annotator.

Furthermore, the ontology and the annotated instance data should be in a form that is easy to use by the

application programmer and efficient to run by the exhibition software. In their work, two major difficulties were encountered during the annotation and implementation process:

- Annotation process posed new demands to the ontology, which lead to changes in the ontology after many annotations were already done. How to manage such changes so that the annotator wouldn't have to redo the annotation work?
- Application programmers pose new demands to the ontology and to the annotations in order to satisfy the demands of the end-user interface. As a result, changes in both the ontology and the annotations were needed.

Waqas Ahmad, Ch Muhammad Shahzad Faisal [5] proposed context based search of Personality Images, which helps user to find out required images efficiently. Images were gathered from Google, selecting 300 pictures of 5 personalities which were manually annotated. Main context for finding images related to personalities include activities like playing, attending meeting etc.

Steps are as Follows

- Data collection details
- Context definition
- Inferencing
- Queries
- Proposed Image Ontology

Noman Hasany, Mohd. Hasan Selamat [6] presented a system that populates hotel related information in the ontology and use a natural language query platform to retrieve the information from a common interface for decision making. User will get information from a single interface based on user selected parameters. This paper provides the detailed construction of the Hotel Ontology using knowledge base of Malaysian hotels.

Hotel Search Model

Hotel Queries

For the selection of a hotel, a user queries for comparative information for decision making. Usually, the users want to seek information with some other associated information e.g. hotels with area and rankings, room rates with facilities in some area etc. If this information is provided from a common interface, not only the user time is saved but they can compare and utilize the information easily and effectively.

Hotel Ontology Design

The hotel ontology termed here as MyHotel ontology is designed for the system to incorporate the hotel concepts and specific Malaysian hotels data in the form of instances and datatype values.

Wrapper Module

To populate the ontology with online hotel data, a wrapper is needed that performs extraction of the data and instances related to selected ontological concepts from the website and then populate the ontology.

Query Conversion

The answer is obtained by converting the query in semantic web query language i.e. SPARQL.

Tuan-Dung CAO, Thanh-Hien PHAN, Anh-Duc NGUYEN [7] presented an ontology based approach for developing STAAR (Semantic Tourist Information Access and Recommending). It helps tourist search information by providing a various semantic search feature in a mobile phone application using web services. In addition they also proposed an algorithm for recommending travel route relevant to both criterions: itinerary length and user interest.

The main architecture of STAAR system is divided into following modules

- **Semantic Data Integration and Management:** This component is responsible for integrating proprietary data of system with proper external data from Linked data repositories such as DBPedia, GeoNames.
- **Web Service:** To support multi-platform tourist guide applications in the role of client in a distributed environment, a set of web services is developed.
- **Client:** STAAR system is currently deployed for both the web and the Android phone. The web-based system that allows searching tourist resources in accordance with the preferences expressed in a semantic profile and suggested itineraries for visitors on the set of locations, while the Android phone-based system allows users to search quickly a system resource with a simple constraint as well as with complex constraints on one or more objects.

The heart of the system is an ontology that was built on RDF++. RDF++ is a extension version of RDF which supports some important properties of OWL and is simpler than OWL.

K.Palaniammal, Dr. M. Indra Devi, Dr.S.Vijayalakshmi [8] in their paper are concerned with the development of the model towards the semantic search and the result which is based on user's priority while searching the tourism domain of interest. Their system makes use of user's profile ex age, user's current status etc to understand his behavior and tells the probability of his highest interest on the type of places he/she wishes to tour.

System Implementations

Building Ontology

Ontologies are the core technology for the Semantic Web. The ontology is used in e-tourism research, is usually build by ontology web language like OWL. OWL supports some features such as cardinality constraints and data types, which are not supported by RDF. This paper made use of tourism ontology for querying upon the desired event, which consists of classes, properties and individuals related with the tourism domain.

Calculate the Conditional Probability

In this paper to generate the probability for a required JAR file of Netica-J is being added to the library. The file system bayesnetwork.dne is also being imported to the project.

Semantic Query

In this paper, SPARQL query is being used to retrieve tourism relevant information from tourism ontology. SPARQL (Simple Protocol And RDF Query Language) is the same as to SQL and used to access more reliable and accurate results.

P. Sheba Alice, A. M. Abirami, A. Askarunisa [9] in their paper proposed a tool enhancing a refined search retrieving only the most relevant links eliminating the other links using semantic web technologies. A user's searched text is stemmed and compared with attributes defined in ontology.

Hannah Bast, Florian Bäurle, Björn Buchhold, Elmar Haussmann [10] in their paper discussed the advantages and shortcomings of full-text search on the one hand and search in ontologies on the other hand. They say that full-text query work well when relevant documents contain the keywords or simple variations of them in a prominent way. Entity oriented queries work well in ontologies. This paper also discussed the challenges while obtaining the facts for the ontologies.

Document-Oriented Queries

This works well as long as (i) the given keywords or variants of them occur in enough of the relevant documents, and (ii) the mentioned prominence of these occurrences is highest for the most relevant documents. For large document collections (as in web search), the number of matching documents is usually beyond what a human can read. Then precision is of primary concern for such queries, not recall.

Entity-Oriented Queries

Consider the query plants with edible leaves. The first problem is as follows. Relevant documents are likely to contain the words edible leaves or variations of them. But there is no reason why they should contain the word plants, or variations of it like plant or botany. Rather, they will contain the name of a particular plant, for example, broccoli. This is exactly the kind of knowledge contained in ontologies.

The second problem is that the sought-for results are not documents and also not passages in documents, but rather a list of entities, plants with a certain property. Worse than that, the information for a single hit could be spread over several documents. For example, for the query plants with edible leaves and native to Europe, the information that a particular plant has edible leaves may be contained in one document, while the information that it is native to Europe may be contained in another document. This is beyond the capabilities of full-text search engines.

GOAL OF RESEARCH

The main goal of this research work is to develop system architecture for semantic web search engine using concept matching over a specific domain that is Hotels.

Search engine will be trained with the knowledge base in the form of Ontology which can be in any specification like RDF, OWL or Turtle.

Various classes of Hotels are prepared on the basis of Location, Ratings, RatePerRoom etc. User will be provided with easy to use interface to enter his query. Searching will be done in the ontology itself after which images of the relevant Hotels will be displayed based on their ranking. Image ranking has been done based on the preferences of users clicking their websites, thus giving a self learning framework to user similar to that of Google image search.

CURRENT SYSTEM ISSUES

The current search system based on semantic web does not takes into consideration the user behaviour while displaying the search results, which is the important requirement in order to make user aware of the most popular as well as reliable hotels images among Internet users. So our semantic search system focus on using the image popularity as well as the rank of image's host Web Site registered with Google.

PROPOSED SYSTEM

Here in this proposed system we design and develop a semantic web architecture that can relieve the users from the overburden of doing a lot of keyword based search before getting the desired result. This system takes the user query in

the form of parameters related to that concept in a user friendly environment. The proposed system architecture has to be divided into 2 different modules: First module takes the user query in the form of concept description and the second module provides a mechanism to display search results in ranked order that is in the preferences observed by user behavior. But for the system to understand user query and give best result, it should be trained first.

SYSTEM ARCHITECTURE

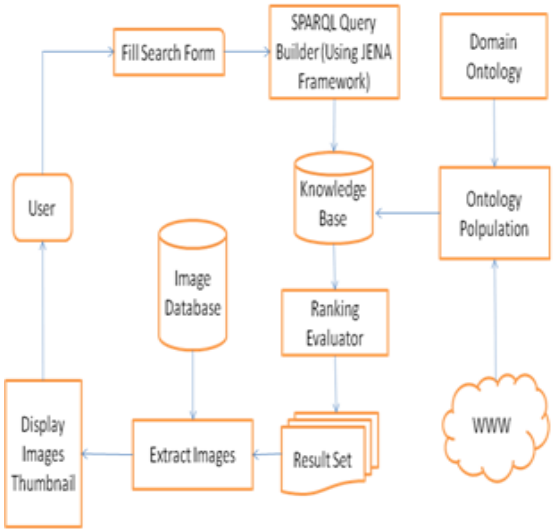

Figure 3: System Architecture

TRAINING THE SYSTEM

The data of various users is collected and organized around ontology of hotels. The system has a large database of images belonging to various categories. These images are passed into an algorithm which extracts various metadata of image such as file type, file size, file dimension, date created on etc.

All the details along with the URL of image file and it s category is stored in a database. The category of an image is identified manually and it can be anything like Location, Ratings, RatePerRoom etc.

CONCLUSIONS

In my research, I have proposed a model implementing a prototype showing the application of semantic web in representing and accessing the information about hotels which will help users in searching and retrieving relevant images from the knowledge base displayed in proper ranked order.

REFERENCES

1. Jiang Huiping, “Information Retreival and Seamtic Web”, 20I0 Technology (ICEIT 2010), 78-1-4244-8035-7/10 2010 IEEE V3-461.
2. WANG Yong-gui1, JIA Zhen2, “Research on Semantic Web Mining” , 2010 International Conference On Computer Design And Appliations (ICCDA 2010), 978-1-4244-7164-5 2010 IEEE.
3. Saman Kamran, Fabio Crestani, “Defining Ontology by Using Users Collaboration on Social Media”, CSCW 2011, March 19–23, 2011, Hangzhou, China. ACM 978-1-4503-0556-3/11/03.
4. Eero Hyv¨onen, Avril Styrman and Samppa Saarela , “Ontology Based Image Retrieval”, University of Helsinki, Department of Computer Science.

5. Waqas Ahmad, Ch Muhammad Shahzad Faisal, “Context Based Image Search”, 978-1-4577 -0657 -8/11/$26.00 © 2011 IEEE .
6. Noman Hasany, Mohd. Hasan Selamat, “Answering User Queries from Hotel Ontology for Decision Making”, 978-1-4577-1481-8/11/$26.00 ©2011 IEEE 123 .
7. Tuan-Dung CAO, Thanh-Hien PHAN, Anh-Duc NGUYEN, “An Ontology based approach to data representation and information search in Smart Tourist Guide System”, 978-0-7695-4567-7/11 $26.00 © 2011 IEEE DOI 10.1109/KSE.2011.33.
8. K.Palaniammal, Dr. M. Indra Devi, Dr.S.Vijayalakshmi, “An Unfangled Approach to Semantic Search for E-Tourism Domain”, 978-1-4673-1601-9/12/$31.00 ©2012 IEEE 130 ICRTIT-2012.
9. P.Sheba Alice, A.M.Abirami, A.Askarunisa, “A Semantic Based Approach to Organize eLearning through efficient Information Retrieval for Interview Preparation”, 978-1-4673-1601-9/12/$31.00 ©2012 IEEE ICRTIT-2012.
10. Hannah Bast, Florian Bäurle, Björn Buchhold, Elmar Haussmann, “A Case for Semantic Full-Text Search”, JIWES ’12 August 12 2012, Portland, OR, USA Copyright 2012 ACM 978-1-4503-1601-9/12/08 ...$15.00.
11. http://jena.apache.org/documentation/ontology/
12. http://answers.semanticweb.com/questions/2863/insert-data-into-ontology-and-how-to-query
13. http://searchengineland.com/semantic-search-what-is-it-how-are-major-search-and-social-engines-use-it-part-1-133160
14. http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html