The Research and Implementation of Semantic Based RDF Tagging and Webpage Searching Web Service

*Jason C. Hung, **Schummi Yang, **Mao-Shuen Chiu and **Timothy K. Shih
*Department of Information Management
Northern Taiwan Institute of Science and Technology
No. 2, Xueyuan Rd., Peitou, 112 Taipei, Taiwan, R.O.C
**Department of Computer Science and Information Engineering
Tamkang University
Tamsui, Taipei Hsien, Taiwan, R.O.C.
E-Mail:692192148@s92.tku.edu.tw

Abstract

In recent years, the network and information technology is more and more ripe we always search for data or files through the Internet. The resources on the World Wide Web are increased day by day. The amount of information exchange and speed of update are also grown. Many methods depend on the match of vocabularies between information request and searched objects. For example, we usually adopt some keywords and Boolean operators to form the query to search the Internet for interested information. Unfortunately, users do not always use proper words and operators to form the query for the search. The result is related or unrelated information is both retrieved .The retrieval method becomes critical for us to get more accurate results. In order to improve the searching process, a RDF-based mechanism integrated with word sense disambiguation technique is proposed to semantically index and retrieve web pages on the World Wide Web. The approach is to describe the resources by RDF(S) metadata and store them. The proposed method also has additional advantage of the ability of further integrated into the Semantic Web Service.

Keywords: Semantic Web, WordNet, RDF, Jena, Web Service.

1. Introduction

A large amount of information is retrieved from the Internet by using search engines. A search engine requires one or more user input keywords to carry out a search, but sometimes the search results do not match expectations from users due to the huge amount of information can be accessed through Internet. Network information technology is unceasingly developed and mature, when we want to seek for some information, we often search the WWW. The resources on the network are increasing day by day, and the information exchange and absorption rate are also multiplying. The search engine, like Google, has three hundred million inquiries and 40 hundred million indexed homepages every day in the World Wide Web. The automated searching usually uses the web crawler, spider, robot (bot) or the agent technology, follows the HTTP hyperlink between websites to search and collect web pages on the Internet.

The W3C **[4]** research team proposed the RDF (Resource Description Framework) standard **[8]**. RDF is based on XML **[9]** and in XML grammar foundation. RDF stipulates the metadata storage structure and related technical standard. Using the RDF language, we can characterize the information in a uniform, exchangeable style. Furthermore, this makes it possible for machines to "understand" the data.

If the search engine searches the Internet depending on the conceptual matches, not just relying on the similar word usage, it would have better search ability to respond the detail inquiry request. Using RDF tagging may provide opportunity for new search methods. But, the majority of semantic search engine will meet the potency problem when searching for information within the massive semantic network. In order to obtain the effective searching results, the network must contain massive related information. But meanwhile, the large-scale network would cause to discover a best way in many processing solutions question way to appear is not easily.

However, the Internet search has some problems. For example, when using nowadays search system, we always utilize keywords to query. But to make the inquiry in the material huge information sea, we often fall into another information sea, even more information sea. Thus we may find more unrelated data than truly needed data, the searching efficiency will thus reduce very much. The cause of this condition is because the

words meaning are confused or the improper resources description or tag. We knew that it is an issue. In general, much of the data in World Wide Web has less correlation defined between each other, and this makes it difficult to search related material from one to another.

We proposed a mechanism to achieve the semantic web service system. It includes the semantic searching method, combines WordNet **[2]** and RDF to search the different characteristic resources from the Internet. The goal is to supply interaction and commutation between heterogeneity resources, to annotate them by RDF and WordNet metadata, and to supply other agent process to inquiry these RDF tagged documents. It achieved semantic web page resources sharing progress for human natural operation and convenience.

Therefore, according to the concept-based WSD system model proposed by Che-Yu Yang (AINA'2005) **[1],** study of this paper will focus on using RDF description along with WordNet **[3]** Synset (synonymy word collection) to describe/tag semantic meaning of web-pages content, as well as using the characteristic – Notation Triples (N3) – to give signs for marking these various resources with other resources. The query language for RDF documents is Jena **[6]** which provides N-Triple to search the words relative. Finally we established the Semantic Web Service, achieved the goal to supply the semantic query and sharing resources.

This paper is organized as follows. In section 1 we give a brief introduction about the research background and our system. We talk about the searching advantage of our idea. Next, in section 2 is related work about the RDF and its query language. In section 3 we discuss about the relationship between RDF and the WordNet. We established a method to add the semantic concept into web pages that can help the process of search and query. In section 4 we integrate the whole system and package it as Web Service. Finally is the conclusions and future work in section 5.

2. Related Work

The Resource Description Framework (RDF) is a W3C Recommendation for the formulation of metadata-description on the World Wide Web. The RDF is a simple model and considered to be the most relevant standard for data representation and exchange on the Semantic Web **[10]**. The RDF Schema (RDFS) extends this standard with the means to specify domain vocabulary and object structures in order to describe and define grammar as well as the announcement the RDF. RDFS looks like a dictionary, it describes each property significance, the characteristic, and the constraint of property value. RDFS may let the person to read and understand each data attribute significance. RDFS defines the class and property to describe the resources content.

Also RDF is for knowledge and metadata representation. And it is as well fitting for representing any data or metadata. RDF may be regarded as one kind of Web knowledge to express the language, or said it is a logical language. RDF has the formalized grammar, the semantic model, the ability to prove inference as well as and theorem of the reliability. The architecture of RDF is based on the Extensible Markup Language (XML). Therefore we may use RDF data model and use the directional characteristic to construct the relation. So we will use RDF to describe metadata.

We integrated the RDF Query Language (RDQL). RDQL is used to query RDF documents language in the tradition of database. RDQL is a typed language for generalized path expressions featuring variables. In order to query the RDF documents, we use the Jena's API architecture to focus on the RDF data model. Jena is Java toolkit for developing semantic web applications based on W3C recommendations for RDF and OWL. It provides an RDF API; ARP, an RDF parser, RDQL, an RDF query language; an OWL API; and rule-based inference for RDFS and OWL. A basic RDF/XML document is created by instantiating one of the model classes and adding at least one statement (N3) to relate them for accessing the metadata or elements attribute.

We also utilized the Sesame **[7]**. Sesame is an open source RDF database with support for RDF Schema inferencing and querying. Sesame is a Java framework for storing, querying and inferencing for RDF and RDFS **[12]**. It can be deployed as a web server or used as a Java library. Features includes that allow persistent storage of RDF data and schema information and subsequent querying of that information.

3. Design and Functional Requirements for the RDF Storage and Query

As the extension of the concept-based word sense disambiguation (WSD) model which was proposed by Che-Yu Yang **[1]**, this paper will focus on using RDF format, combined with WordNet ontology, to indexes/tags web pages with concepts (actually the synonymy sets, usually called "synsets"). We also use the characteristic – Notation Triples (N3) to mark these various resources with other resources constructions.

When users want to query something by the keywords, through the WSD module **[1],[10]** and Google API **[5],** we can filter out semantically unrelated web pages (miss-retrieved) and leave only conceptually matched web pages. Next, we use the synonymy sets (synsets) in WordNet with the N3 notation in RDF to annotate the keywords on the web pages with their own concepts/senses in the context. When users use keywords to search the Internet, the keywords will be disambiguated by the WSD module and assigned/tagged with each WordNet's synset-id to each keyword

according to their meanings/senses in the context. Not only the keywords in the user query, but also the keywords in the conceptual matched web pages after disambiguation are tagged/annotated with synset-ids. That's actually the semantic mapping between keywords in web pages and senses/concepts in WordNet, as shown in the figure 1.

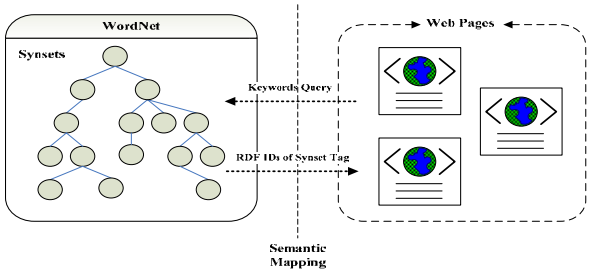

Figure 1. Semantic mapping between WordNet and web pages

We can use N3 notation (subject, predicate and object) of RDF to achieve the above idea. Using N3 notation we can establish the relation between the concepts/senses in Wordnet and keywords in web pages to archive the RDF/RDF(S) characteristics. And we construct the N3 notation by the synset-id and URI (rdf:alt), as shown in figure 2. It represents the properties and attributes of the resources type. The word (book) is oriented towards the relative web pages (URIs) and the arrow point means to the property each other(wn:). The "wn:" represents the Wordnet RDF schema.

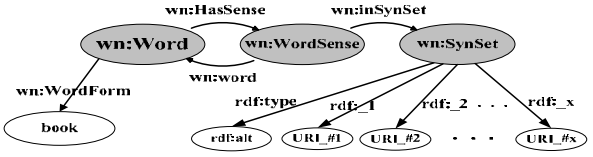

Figure 2. The semantic relations of the Wordnet in RDF

4. System Architecture and Implementation

We employ the glossaries defined in WordNet and in WordNet Schema, and use the synset-id of glossary itself to establish the connective model, and we also take shape representation of RDF Model. Because the webpage may consist of many different glossaries, we can combine these different glossaries with different URI of WordNet synsets. Finally they will form a hierarchical heterology structure RDF Model. We treat RDF as the reference structure of synset-ids. According to the different webpage resources characteristic we construct RDF Model Sets. The RDQL provides the ability to query the RDF annotated documents.

Before tagging documents with the RDF notation, we must make the index of them first. After parsing the RDF documents (the N-Triples part) and collecting correlative data to store in our database, we establish the entity-relation table to store the transformation results. The method is to treat each N3 (Subject, Predicate and Object) as the same identical unit when storing them. (This method offers more efficiency when we want to do concept searching and index catalog of contents service). This is due to the need of reference and index of the resources service in the future. The architecture diagram is as figure 3.

So, when users query with the keyword(s), the WSD algorithm will semantically filter out the unrelated web pages and leaves only semantically matched ones that are returned from the Internet search request from Google API. And we tag the matched keywords in the web pages with synset-ids when carry out semantic mapping between web pages and Wordnet. Then the system will establish the related property sets which includes the attributes of the subject, predicate and object. They are the classes of words named by URIs. To reference the RDF/RDFS format storage proposed in our system. Finally we package the whole procedure into a Semantic Web Service.

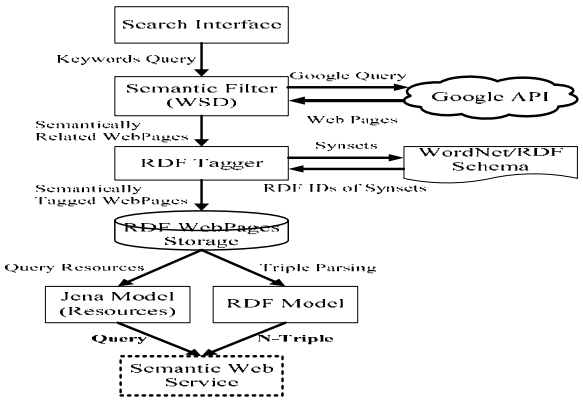

Figure 3. The system architecture

Jena is the standard inquiry language of RDF and RDFS. It provides the diverse searching measures, including files or the website content. Also it can supply to limit the scope of the searching (use "WHERE" Clause), logical determining and filtering ("AND" Clause, "USING" Clause).The Jena development platform is based on Java Framework. It's the application program interface (API) which can construct the query layer of Semantic Web architecture.

We utilized the "Sesame", which supports Jena query language, to integrate the RDQL into our system. And we use "mysql" as the storage database, which includes the XML and RDF documents.

When user puts a query, the system will access the repository (Sesame) to search the RDF documents by Jena API in our database (mysql). According to the properties of N3 we know that the characteristics of the

subject, predicate and object. Figure 4 is our system flow chart.

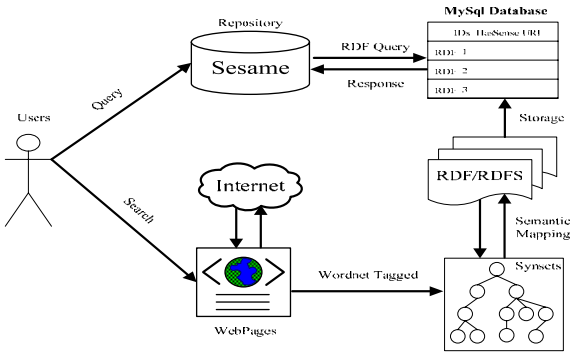

Figure 4. The system flow chart

We construct the Web Service system by Java Web Services Developer Pack (Java WSDP) **[12]**. The environment of our implementation is the Apache Jakarta Tomcat (Apache SOAP 2.0, Java servlets 2.2 API standard), and we installed the Apache Xerces XML Parser 1.2.3. It supplies the XML grammar parsing and supports to develop the most XML standard. Figure 5 is our system interface, as follows.

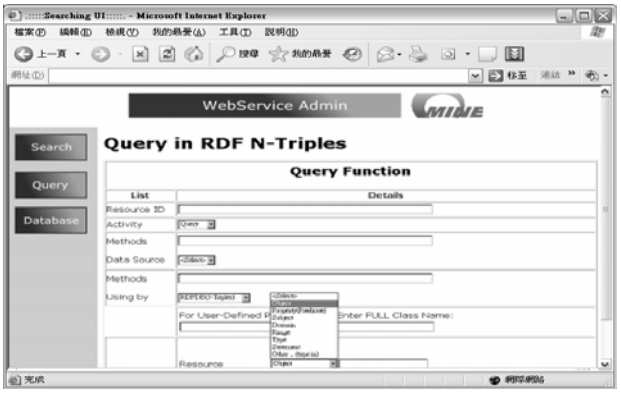

Figure 5. Web Service system

We proposed the semantic Web Service system, including the search and query method to the different resources by the attribute characteristic in WordNet and RDF. Web Service can communicate with different users. Our system achieved the resource sharing, exchanging and communication as well. The goal is the resource sharing to achieve both the human and machine operation convenience.

5. Conclusions and Future Work

The system based on semantic will mark up from the different resources. It offers the ability to search and query the resources on the World Wide Web, finally packs into Web Service system to improve systematic practicability further. We propose the system to improve on searching more semantic and sharing resources.

We will discuss in the study on Ontology in the future. The metadata is packaged up Ontology to integrate relative object. In addition, the OWL (Web Ontology Language) is a component of the Semantic Web framework. OWL is built upon RDF and RDF Schema. OWL also adds more vocabulary for describing properties and classes. It's intended to provide a language that can be used to describe the classes and relations between them that are inherent in Web documents and application.

References

[1]. Jason C. Hung, Ching-Sheng Wang, Che-Yu Yang, Mao-Shuen Chiu, George Yee, "Applying Word Sense Disambiguation to Question Answering System for E-Learning", International Conference on Advanced Information Networking and Applications (AINA 2005).

[2]. George A. Miller, "WordNet: A Lexical Database," Comm. ACM, Vol. 38, No. 11, 1993, pp. 39-41.

[3]. Leacock and Chodorow. 1998. Combining local context and WordNet similarity for wordsense identification. In Fellbaum 1998, 265-283.

[4]. W3C, World Wide Web Consortium, http://www.w3c.org/

[5]. Google API, http://www.google.com.tw/apis/

[6]. Jena, http://jena.sourceforge.net/

[7]. Sesame, http://openrdf.ord/

[8]. Tim Berners-Lee, 2000. "Primer: Getting into RDF and Semantic Web using N3".

[9]. M.Klein, "XML, RDF, and Relatives," IEEE Intelligent Systems, vol. 16, no.2, Mar/Apr 2001, pp. 26-28.

[10]. Jiangsheng Yu,"WSD and Closed Semantic Constraint" supported by National Foundation of Natural Science (Research on Chinese Information Extraction) No. 69483003 and Project 985 in Peking University.

[11]. Boris Katz Jimmy Lin, "Annotating the Semantic Web Using Natural Language, " In Proceedings of the 2nd Workshop on NLP and XML (NLPXML 2002) at COLING 2002, September 2002, Taipei, Taiwan.

[12]. Java Web Service Developer Pack (Java WSDP) http://java.sun.com