Semantic Web Information Retrieval Based on the Wordnet

[1]Che-Yu Yang, [2]Shih-Jung Wu
[1]*Dept of Information Management, China University of Technology, Taipei, Taiwan*
Email: cyyang@cute.edu.tw
[2]*Dept of Innovative Information and Technology, Tamkang University, I-lan County, Taiwan*

Abstract

Most of the existing textual information retrieval approaches depend on a lexical match between words in user's requests and words in target objects. Typically only objects that contain one or more common words with those in the user's query are returned as relevant. This lexical based retrieval model is far from ideal. In this research an approach to semantic based information retrieval of semantically annotated documents is presented. The approach operates based on: (i).natural language understanding, (ii).the Wordnet ontology, and (iii).the Semantic web standards. Not only the information is annotated and searched on a semantic basis, but also the retrieval process can be enhanced by the use of rich vocabulary knowledge in the ontology.

Keywords: *Semantic Web, Semantic Information Retrieval, Ontology.*

1. Introduction

Most existing information retrieval systems are based, either directly or indirectly, on models of the traditional information retrieval (IR) system. These retrieval models specify how to create representation for textual documents, and how these representations and information needs should be compared with each others in order to estimate the likelihood that a document will be judged relevant. The estimates of the relevance of documents to a given query are the basis for the document rankings that are now a familiar part of IR systems. Examples of classic models include the probabilistic or Bayes classifier model [1,2], and the vector space model [3]. Many others [4,5,6,7] have been proposed and are being used.

However, these lexical based retrieval model is far from ideal ─ many objects relevant to user query are missed, and many unrelated objects are retrieved. Some researches show that fundamental characteristics of human verbal behavior result in these retrieval difficulties [8]. Because of the tremendous variety in the vocabulary can be used to describe the same meaning or concept (*synonymy*), people will often use different words from the author or indexer of the information, and relevant materials will therefore be missed. On the other hand, since a single word often has more than one meaning (*polysemy*), irrelevant materials will often be retrieved.

For textual retrieval systems that utilize automatic indexing/searching techniques to create and compare text representatives from natural language to achieve better performance, they must deal with the problems of polysemy and synonymy. Polysemy, a single word form having more than one meaning, decreases retrieval precision by false matches. While synonymy, multiple words having the same meaning, decreases the retrieval recall by missing true conceptual matches.

This research tries to overcome these problems by retrieving textual information via word's meanings, rather than the word's lexical forms. In a previous work [9], an automated approach to extract semantics from textual data and annotate these semantics to the data is proposed. Word sense disambiguation (WSD) technique is used to identify the concepts in context against Wordnet ontology. Then the Resource Definition Framework (RDF) is used to annotate the semantics. I this fashion, much of the existing data can be processed and brought to the Semantic Web. Further in this research, an approach to retrieve data via Wordnet ontology and semantic annotations is proposed. Also, making inference from the ontology to help searching relative data is discussed. Finally a framework to integrate and automate these processes is demonstrated.

2. Related works

In the emerging Semantic Web, search, information interpretation and aggregation can be manipulated by ontology-based semantic annotation. Reference [10] examines the semantic annotation, identify a number of requirements, and review the current generation of semantic annotation systems. It shows that, while there is still some way to go before semantic annotation tools will be able to address fully all the knowledge management needs, research in the area is active and making good progress.

Reference [11], the researchers investigate the definition of an ontology-based IR model, oriented to the exploitation of domain Knowledge Bases to support semantic search capabilities in large document repositories. Another research [12] focuses on a holistic architecture for semantic annotation, indexing, and retrieval of documents with regard to extensive semantic repositories. A system called KIM is proposed, which is a semantically enhanced information extraction system provides automatic semantic annotation with references to classes in the ontology and to instances. Reference [13] a novel metric to measure the semantic relatedness between words is proposed. The approach is based on ontologies represented using a general knowledge base for dynamically building a semantic network. Then obtain an efficient strategy to rank digital documents from the Internet according to the user's interest domain.

Reference [14] an ontology-based user model, called user ontology, for providing personalized information service in the Semantic Web is presented. The proposed user ontology model utilizes concepts, taxonomic relations, and non-taxonomic relations in a given domain ontology to capture the users' interests. The proposed user ontology model with the spreading activation based inferencing procedure has been incorporated into a semantic search engine, called OntoSearch, to provide personalized document retrieval services.

3. Wordnet for semantic web

In [9], the word sense disambiguation technique is utilized to automatically discover the underlying semantic information of the textual data, based on the popular open-domain vocabulary ontology − the Wordnet. This semantic information is then annotated to the documents in RDF that conforms to the Semantic Web standards for future reuse. The whole semantic annotating process is shown in figure 1.

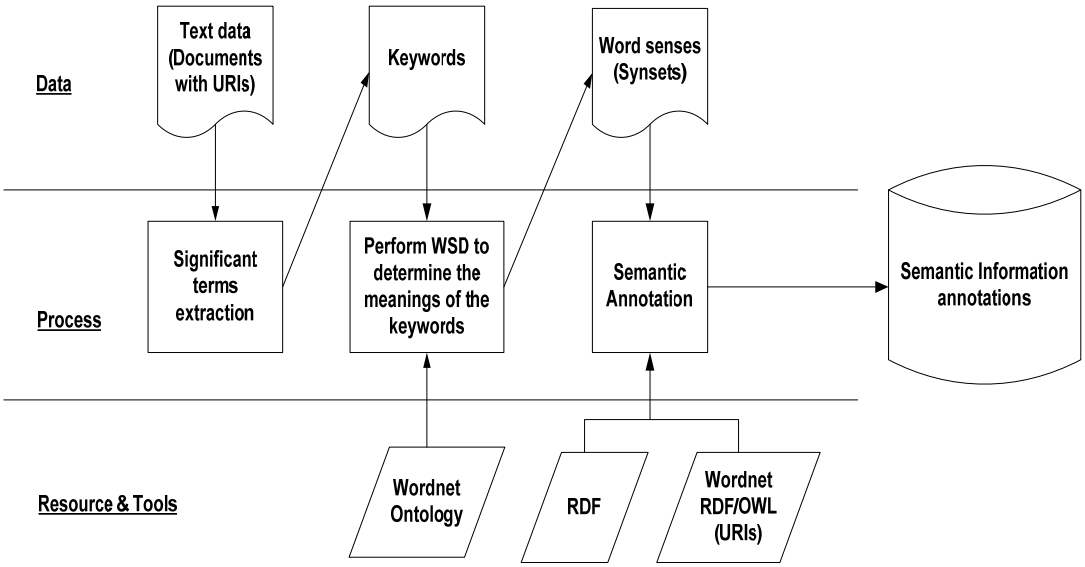

Figure 1. The process of annotating semantics to the textual data.

3.1. WordNet in RDF/OWL representation

Wordnet [15] is a machine-readable dictionary developed by George A. Miller et al. at Princeton University. In the RDF/OWL representation of WordNet, the WordNet schema has three main classes:

Synset, WordSense and Word, and their subclasses. Each instance of Synset, WordSense and Word class is assigned with one distinct URI. The pattern for the instance of a Synset is:
wn20 instances: + synset- + %lexform%- + %type%- + %sensenr%, where

- *%lexform%* is the lexical form of the first WordSense of the Synset.
- *%type%* is one of noun, verb, adjective, adjective satellite and adverb.
- *%sensenr%* is the number of the WordSense that is contained in the synset.

For example, the following URI represents a NounSynset that contains a WordSense which is the second sense of the word "bank":
http://www.w3.org/2006/03/wn/wn20/instances/synset-bank-noun-2

With this RDF/OWL representation, Wordnet becomes an available ontology to the Semantic Web tools, and the URIs for synsets or word senses can be made associated with text data acting as semantic annotations.

3.2. Finding the semantics of the data

Before we can manipulate the textual data with vocabulary ontology at the semantic level, the conceptual semantics of the text data has to be understood first. Word Sense Disambiguation (WSD) [16,17,18,19,20] is the task of figuring out the intended meaning of a word when used in a sentence. It involves labeling every word in a text with a tag from a pre-specified set of tag possibilities for each word by using features of the context and other information.

The WSD can be utilized to figure out the semantics of the text data - that is to create the mapping between a word and a Wordnet synset. All human languages have words that bear different meanings when appear in different contexts, such words with multiple meanings are potentially "*ambiguous*". The polysemy issue can be handled by assigning different senses of a word different concept identifiers, whereas synonymy can be handled by assigning the same concept identifier to synonym words. Accurate word sense disambiguation can lead to better results for information retrieval.

3.3. Semantic annotation to the data with RDF

The Resource Description Framework (RDF) is a language for representing the information about resources in the World Wide Web. In RDF, the structure of a statement is a collection of "triples" consisting of a "subject", a "predicate" and an "object", where:

- Subject: the resource the statement describes
- Predicate: a specific property of the resource the statement describes
- Object: the value of this property for the resource the statement describes

Thus a document, such as a PDF file, a MS Word file, a Web page, a text file, can correspond to a "subject"; a word sense (synset's URI) in Wordnet can correspond to an "object". As to the "predicate" which indicates the relationship between "a document" and "a synset", the term "subject" in Dublin Core[21] is used as the "predicate" in a RDF statement. For example, to assert the statement: "http://about.com/loan.html"" mentions to the concepts "bank" (the 2nd sense of the word bank - financial institution), "money" (the 1st sense of the word money - medium of exchange) and "check" (the 1st sense of the word check - a written order directing a bank to pay money), the RDF graph for the assertion would look like figure 2.

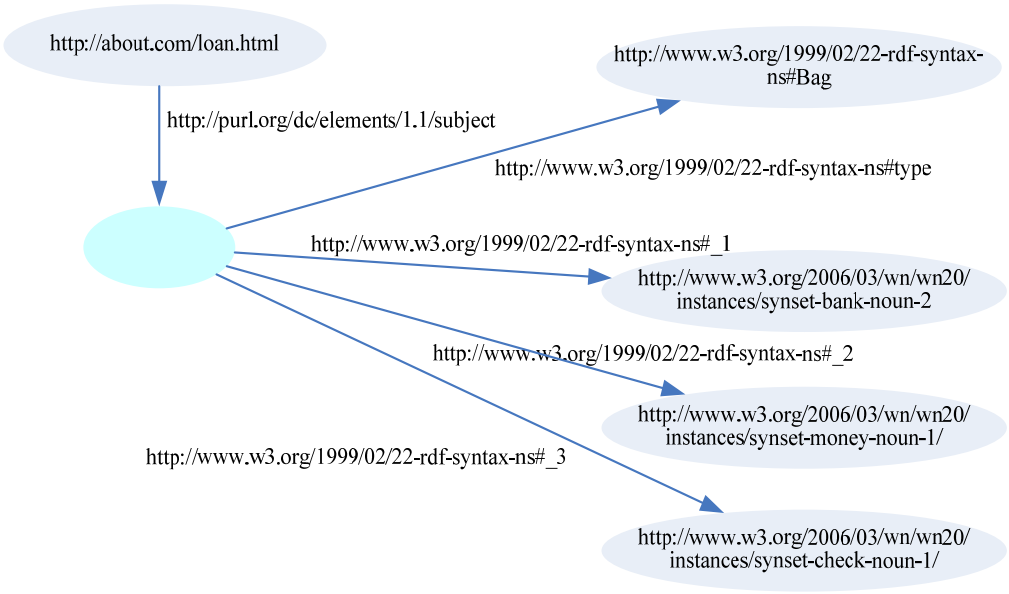

Figure 2. An example RDF graph indicating a document refers to three concepts.

In this way a document (text) is just like a bag filled with "concepts" in it. Furthermore machines can know what a document is about and can access to the document in a way conforming to the Semantic Web standards.

4. Use the semantic annotation to retrieve data

In this section, how to facilitate the retrieval using the semantic annotations of the data described above is investigated. Most of the modern search engines operate based on the traditional information retrieval techniques. They directly and merely deal with the strings (word forms) in the text, and do not take care of the semantic annotations in the documents (they often choose to ignore or treat these annotations as ordinary context). Therefore these techniques are far from sufficient for dealing with the Semantic Web documents which consist of rich semantic annotations such as RDF triples or RDF statements.

One possible approach to utilize the semantic annotations to search data is through the domain knowledge encoded in the ontology that describes the data. First, the semantics of user's information needs should be identified in the ontology then the searching for textual documents is to find the instances of these concepts through the semantic annotations to the documents.

As the semantic information can be annotated to the documents using the RDF graph data model, some corresponding method to query for the information in the RDF graphs in needed. SPARQL[22] is a query language and a protocol for accessing RDF designed by the W3C. As a query language, SPARQL queries the information held in the RDF graphs. It takes the description of what the application wants in the form of a query, and returns that information in the form of a set of bindings or a RDF graph. Used with a common protocol (such as HTTP and SOAP), applications can access and combine information from across the Web.

The SPARQL query language is based on matching graph patterns. The simplest graph pattern is the triple pattern, which is like an RDF triple, but with the possibility of a variable instead of an RDF term in the subject, predicate or object positions. Combining triple gives a basic graph pattern, where an exact match to a graph is needed to fulfill a pattern. For example, figure 3 is a RDF graph that encodes the annotation "The book1's title is SPARQL Tutorial".

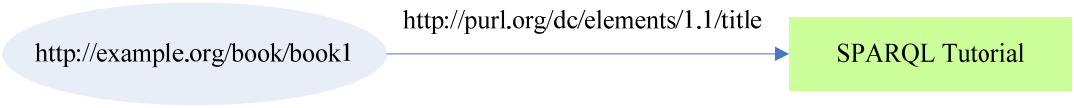

Figure 3. An example of RDF graph.

To find the title of a book from the RDF graph, the SPARQL query is as follows:

```
SELECT ?title
WHERE
{
  http://example.org/book/book1
  http://purl.org/dc/elements/1.1/title   ?title
}
```

The query consists of two parts: the SELECT clause and the WHERE clause. The SELECT clause identifies the variables to appear in the query results, and the WHERE clause has one (or more) triple pattern. The output of the query would be "SPARQL Tutorial" in this example.

If we have the following RDF annotations that describe a documents "http://about.com/loan.html" contains three concepts "bank", "money" and "check":

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix  dc:   <http://purl.org/dc/elements/1.1/>
@prefix  wn:   <www.w3.org/2006/03/wn/wn20/instance/>
@prefix  ab:   <http://about.com/>
ab:loan.html    dc:subject      _:z.
_:z             rdf:type        rdf:Bag.
_:z             rdf:_1          wn:synset-bank-noun-2.
_:z             rdf:_2          wn:synset-money-noun-1.
_:z             rdf:_3          wn:synset-check-noun-1.
```

RDF also provides the XML syntax for writing down and exchanging RDF graphs, called RDF/XML. Unlike triples, RDF/XML is the normative syntax for writing RDF. Thus the same annotation as figure 2 can be written in RDF/XML as follows:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
  <rdf:Description rdf:about="http://about.com/loan.html">
    <dc:subject>
      <rdf:Bag>
        <rdf:li rdf:resource="www.w3.org/2006/03/wn/wn20/instance/synset-bank-noun-2"/>
        <rdf:li rdf:resource="www.w3.org/2006/03/wn/wn20/instance/synset-money-noun-1"/>
        <rdf:li rdf:resource="www.w3.org/2006/03/wn/wn20/instance/synset-check-noun-1"/>
      </rdf:Bag>
    </ dc:subject >
  </rdf:Description>
</rdf:RDF>
```

To query which concepts does the document "ab:loan.html" refer to, may look like:

```
PREFIX  dc:   <http://purl.org/dc/elements/1.1/>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
SELECT  ?concept  WHERE
{
  <http://about.com/loan.html>  dc:subject  ?x.
  ?x  a  rdf:Bag.
  ?x  rdfs:member  ?concept
}
```

On the other hand, to query which documents refer to the concept "wn:synset-bank-noun-2" may look like:

```
PREFIX  dc:   <http://purl.org/dc/elements/1.1/>
```

```
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
SELECT  ?doc  WHERE
{
  ?doc  dc:subject  ?x.
  ?x  a  rdf:Bag.
  ?x  rdfs:member  wn:synset-bank-noun-2
}
```

The SPARQL language specifies four different query variations for different purposes, as listed in table 1.

Table 1. The four different query variations in SPARQL language.

Query type	Purpose
SELECT query	Used to extract raw values from a SPARQL endpoint, the results are returned in a table format.
CONSTRUCT query	Used to extract information from the SPARQL endpoint and transform the results into valid RDF.
ASK query	Used to provide a simple True/False result for a query on a SPARQL endpoint.
DESCRIBE query	Used to extract an RDF graph from the SPARQL endpoint, the contents of which is left to the endpoint to decide based on what the maintainer deems as useful information.

Each of these query forms takes a WHERE block to restrict the query although in the case of the DESCRIBE query the WHERE is optional.

5. Improve searching with ontology inference

When search for data with the ontology, it is feasible to extend or expand the search to the relative items to the given one. This process can be iterative: to find more items from previous results. For instance, the search inference may take a concept of vocabulary as an input, use the semantic relations of vocabulary encoded in ontology to acquire some relative concepts, and then use these relative concepts to do further searching.

The Wordnet ontology is a kind of semantic net that consists of nodes (synsets) that represent unique concepts, and nodes are in term connected to each others through semantic relations. These nodes and semantic relations used for exploring concepts from one to others during the search, e.g. search for the data contains opposite concept to the original one. Table 2 lists some major semantic relations between concepts.

Table 2. Major semantic relations between concepts defined in Wordnet.

Semantic Relation	Meaning	Example
Synonymy	*X* is similar to *f(X)*	homo, man, human being, human
Hypernym	*X* is a kind of *f(X)*	Apple *is a kind of* fruit
Hyponym	*f(X)* is a kind of *X*	Zebra *is a kind of* Horse
Holonym	*X* is a part/member of *f(X)*	Wheel *is a part of* a car
Meronym	*X* has part/member *f(X)*	Table *has part* leg
Antonym	*f(X)* is the opposite of *X*	Wet *is the opposite of* dry

Take the word “car” as an example. The word “car” may represent five different concepts, one of them is: “a motor vehicle with four wheels; usually propelled by an internal combustion engine”. The RDF/OWL representation of this “car” concept is in the form of URI as:

http://www.w3.org/2006/03/wn/wn20/instances/synset-car-noun-1

If a user is searching for documents contain this “car” concept, without the rich lingual semantic information in Wordnet ontology, the search would be comparatively narrow limiting the scope of results. On the other hand, with the rich information in the Wordnet ontology, numbers of concepts related to the original concept “car” can be found through particular semantic relations. If a user needs the data about particular instances of “car”, then the hyponym relation can be adopted for inference, then the related concepts such as “compact”, “coupe” and ”sedan” are then achieved. These particular concepts can then be use to query for the data contains them. Or if the user wants to find the data talking about things that have similar idea to “car”, then the relations “sister term” can be used. Such a process to reach semantically related concepts using Wordnet ontology is shown as figure 4.

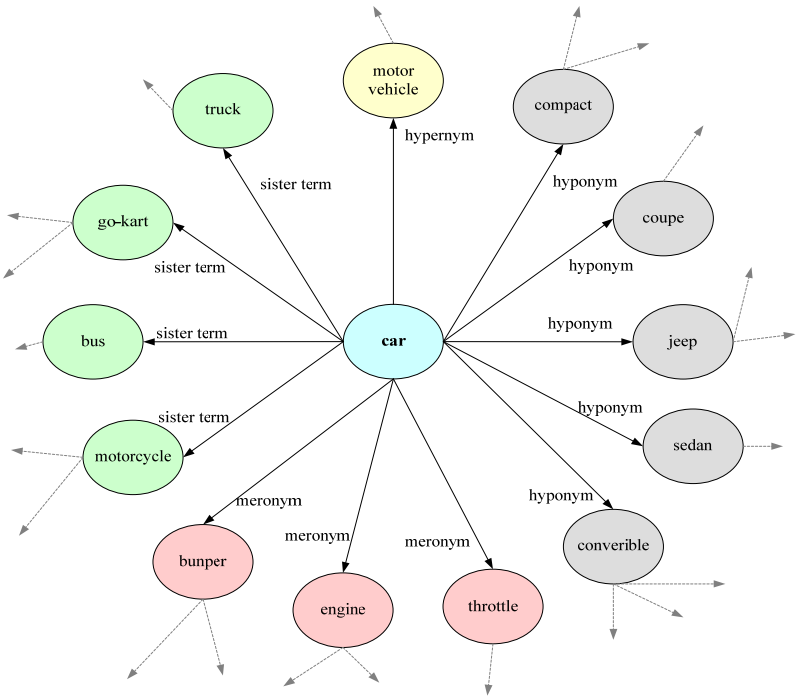

Figure 4. Related concepts can be found through the semantic relations in Wordnet ontology.

In the form of Wordnet RDF/URI, through various semantic relations, the concepts related to the concept “*http://www.w3.org/2006/03/wn/wn20/instances/synset-car-noun-1*“ can be found, as shown in table 3. (With the prefix wn: http://www.w3.org/2006/03/wn/wn20/instances/)

Table 3. Concepts related to the concept “car” can be found through semantic relations.

Semantic relation	Related concepts
Hyponym	wn:synset-compact-noun-3
	wn:synset-coupe-noun-1
	wn:synset-jeep-noun-1
	wn:synset-sedan-noun-1
	wn:synset-convertible-noun-1
Meronym	wn:synset-throttle-noun-2
	wn:synset-engine-noun-1
	wn:synset-bumper-noun-2
Hypernym	wn:synset- motor vehicle-noun-1
Sister term	wn:synset-motorcycle-noun-1
	wn:synset-bus-noun-1
	wn:synset-go-kart-noun-1
	wn:synset-truck-noun-1
...	…

In the retrieval progress, documents can be retrieved according to the concepts contained in their content. Besides, through the semantic relations in the ontology, related concepts can be achieved so that related documents can be further retrieved if needed. The flow of document retrieval based on their semantic annotations is as figure 5.

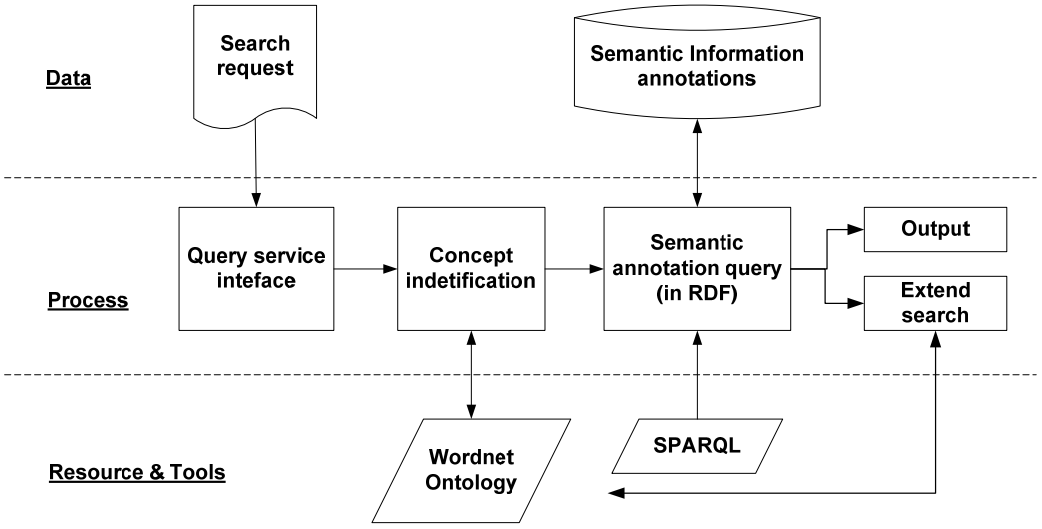

Figure 5. The flow of textual data retrieval using semantic annotations and Wordnet ontology.

6. Conclusions

The realization of the Semantic Web requires the widespread availability of semantic annotations based on ontologies for existing and new documents on the Web. Semantic annotations are to tag ontology class instance to data. The fully automatic creation of semantic annotations is an emerging issue. The retrieval approach that can effectively utilize the semantically annotated metadata to facilitate the search is definitely needed.

This work proposes information retrieval model for textual data in the way to:

1. Extract semantic information from the textual data using WSD technique;
2. Annotate data with the semantics in the fashion conforms to the RDF standard;
3. Retrieve data through queries to the semantic annotations with SPARQL;
4. Make inference from the ontology to enhance the Retrieval.

This IR model will facilitate the migration from today's Web to the future's Semantic Web

7. References

[1] S. Robertson, S. Jones, "The probability ranking principle in information retrieval," Journal of Documentation, vol. 33, pp. 294-304, 1977.
[2] C. Van Rijsbergen, Information Retrieval, Butterworths, London, 1979.
[3] G. Salton, A. Wong, C. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, pp. 613-620, 1975.
[4] C. Van Rijsbergen, "Anon-classical logic for information retrieval," Computer Journal, vol. 29, pp. 481-485, 1986.
[5] S. Deerwester et al., "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, pp. 391-407, 1990.
[6] N. Fuhr, C. Buckley, "A probabilistic learning approach for document indexing," ACM Transactions on Information Systems, vol. 9(3), pp. 223-248, 1991.
[7] H. Turtle, W. Croft, "Evaluation of an inference network-based retrieval model," ACM Transactions on Information Systems, vol. 9(3), pp. 187-222, 1991.
[8] G. Furnas, T. Landauer, L. Gome, S. Dumais, "The Vocabulary Problem in Human-Systems

Communication," Communications of the ACM, vol. 30(11), pp. 964-971, 1987.
[9] Che-Yu Yang, Hua-Yi Lin, "*Semantic Annotation for the Web of Data ─ An Ontology and RDF based Automated Approach*", Journal of Convergence Information Technology (JCIT), Special Issue on Social Networks Application for Decision Support, 2011.
[10] V. Uren et al., "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 4(1), pp. 14-28, 2006.
[11] M. Fernández et al., "Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web, in press, 2011.
[12] A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff, "Semantic annotation, indexing, and retrieval," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 2(1), pp. 49-79, 2004.
[13] A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the Web," ACM Transactions on Internet Technology (TOIT), vol. 9(3), article no. 10, 2009.
[14] Xing Jiang, A.H. Tan, "Learning and inferencing in user ontology for personalized Semantic Web search," Information Sciences, vol. 179(16), pp. 2794-2808, 2009.
[15] G. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, vol. 38, No. 11, pp. 39-41, 1995.
[16] A. Novischi, M. Srikanth, A. Bennett, "Lcc-wsd: System description for English coarse grained all words task at semeval 2007," in Proc. of the 4th International Workshop on Semantic Evaluations, pp. 223-226, Prague, Czech Republic, 2007.
[17] L. M'arquez, G. Escudero, D. Martinez, G. Rigau, "Supervised corpus-based methods for WSD," Word Sense Disambiguation: Algorithms and Applications, Eds. Springer, New York, NY, pp. 167-216, 2007.
[18] R. Navigli, P. Velardi, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(7), pp. 1075-1088, 2005.
[19] C.Y. Yang, "Word sense disambiguation using semantic relatedness measurement", Journal of Zhejiang University SCIENCE A, vol. 7(10), pp. 1609-1625, 2006.
[20] Ming Che Lee, Hui Hui Chen, Yu Sheng Li, "FCA based Concept Constructing and Similarity Measurement Algorithms", IJACT, Vol. 3, No. 1, pp. 97 - 105, 2011
[21] Dublin Core Metadata Basics, available at: http://dublincore.org/metadata-basics/
[22] E. Prud'hommeaux et al., "SPARQL Query Language for RDF," available at: http://www.w3.org/TR/rdf-sparql-query/, 2008.