Ontology based Books Information Retrieval using SPARQL

Naveen Kumar
Ambedkar Institute of Advanced Communication Technologies & Research

ABSTRACT
The huge number of available documents on the Web makes finding of relevant ones a challenging task. In the current web, it is not possible to manage information resources manually and intelligently. And most of the users prefer more accurate information. Semantic retrieval becomes the emerging topic in information retrieval on the Semantic web. To overcome the limitation of current web, The Semantic web, metadata and ontology will play important roles for advanced information retrieval system. In this paper we provide a mechanism for semantic retrieval of books information in WWW. Firstly we build domain ontology of books, and then we present the semantic retrieval system of books information using SPARQL query language which retrieves the information according to the relation of "synonymy of", "kind of", "part of" between the books concepts.

Keywords
Semantic retrieval, books information, semantic web, information retrieval, SPARQL, semantic web.

1. INTRODUCTION
A major problem with searching on the Web today is that data available on the Web has little semantic organization beyond simple structural arrangement of text, declared keywords, titles, and abstracts [1]. The advent of computers made possible to store large amounts of information. The process of retrieving useful information from huge collections of data is a major concern. The field of Information Retrieval (IR) was born out of this necessity. The field has grown considerably from simple keyword based search to semantically enhanced ontology driven information retrieval systems. All the traditional Information retrieval techniques are full text retrieval technique, which accomplish retrieval by content, index and keywords has some limitation or problems given below.

- In some cases the user would not be able to explicitly express the retrieval need through some simple keywords; therefore the quality of retrieval cannot satisfy the user needs.
- Unable to represent the essential content of information due to the meaning of source information is not expose.
- Provides large numbers of results blindly that users haven't enough time and energy to deal with it.
- The limitation holds prominent in cases such as: Polysemy: "java" as language vs. "java" as coffee. Synonymy: "movies" vs. "films". This causes ambiguity during query processing and leads to unwanted results.

A semantically enhanced information retrieval overcomes the limitations of keyword based search. A keyword based information retrieval method has limited capabilities to grasp and exploit the conceptualizations involved in user needs and content meanings. On the other hand, a semantic search focuses on meanings rather than literal strings. The focus is on semantics leading to better and accurate results. A semantic method uses ontology to overcome the limitations of keyword-based information retrieval [2]. Ontology, with the excellent concept hierarchy and appropriately supporting for logic reasoning, is used widely in information retrieval, especially in the semantic retrieval. For the emphasizing of matching based on knowledge and semantic, semantic retrieval has good performance in recall and precision [3]-[4]. At present, the famous programs, which introduced ontology, include (Onto) Agent [5], Ontobroker [6] and SKC (Scalable Knowledge Composition) [7].

With the development of educational system, the significance of Books information becomes more and more obvious and practical. In the open WWW, huge amount of books information is available, so there is a requirement of making a relevant search according to the users need. The structure of this paper include concept of domain ontology model in second section, we build books ontology model in third section and then in fourth section we proposed a system for semantic retrieval of books information using SPARQL query language and finally we make a conclusion for this paper.

2. DOMAIN ONTOLOGY
An Ontology is a "formal, explicit specification of a shared conceptualization of a domain of interest". Thus, an Ontology is the attempt to express an exhaustive conceptual scheme within a given domain, typically a hierarchical data structure containing all the relevant entities, their relations and the rules within that domain [8].

An Ontology is a 5-tuples $O = (C, P, R, I, A)$, where:

- C represents classes or domain concepts and can be arranged in inheritance hierarchies. They should give the specific definition of concepts both in syntax and semantics level.
- P is a set of concept properties.
- R is a set of binary semantic relations defined between concepts {one-to-one, one-to-many, many-to-many} is the set of relation type.

A set of basic relations is defined as :
R= {synonym-of, kind-of, part-of, instance-of}, which have the following interpretations;

- Part-of relation depicts relation of part and integrity between two concepts.
- kind-of relation is represented by characteristics of inheritance relationship of two concepts.
- Instance-of relation describes inclusion relationship between a concept and its subordinate instance.
- Synonym-of relation depicts equivalence relation between concepts.

- A is a set of axioms. An axiom is a real fact or reasoning rules.

3. BOOKS ONTOLOGY

Ontology construction is carried out using Protégé 4.1 tool shown in figure 1 [9], Protégé is an ontology editor tool. Generally we think that the five rules proposed by Gruber in 1995 are influential:

- Explicitly and objectivity. Terms should be defined by natural language using ontology explicitly and objectivity.
- Integrity. The definition is complete and be able to express the meaning of specific terms.
- Consistency. There are no conflicts between the conclusions producing from knowledge reasoning and the meaning of the term itself.
- Maximum one-way scalability. It is not necessary to amend existing terms when we add the generic or specific terms to ontology.
- The fewest constraints. The constraint conditions of modeling should be limited as little as possible.

Books ontology must cover a wide range of information and must have information granularity for searching information effectively. So we create Books ontology shown in Figure 2, which includes the basic concepts about Books and the semantic relationships.

There are mainly four relationships between concepts including inheritance relation (i.e. kind of), instance relation, part-whole relation and synonymy relation. Inheritance relation shows inclusion relation between the concepts, namely, the sub-concept is a kind of the parent concept, such as "Java" is a kind of "Technical book", in which "Technical book" is the parent concept, and "Java" is the sub-class. Instance relation is a specific existence of concepts, such as "IT department book" is an instance of "Computer book". Part-whole relation expresses that a concept is part of another concept, such as "lab manual" is part of "electronics department book". There is rich semantic information described explicitly in ontology. We can know from Figure 1 that Java is a CSE department book of computer books, and people may call it by the name "OAK".

RDF (Resource Description Framework) and RDFS (RDF Schema) is a data model and support mechanism for representing meta-data of schemas [10], and is an ontology representation language. RDF is an XML-based standard for describing resources that exist on the Web, intranets, and extranets. RDFS is used to create vocabularies that describe groups of related RDF resources and the relationships between those resources.

In RDF/RDFS the concepts of domain ontology are represented as follows:

```
<rdfs:Class rdf:ID="Books"></rdfs:Class>
<rdfs:Class rdf:ID="Category"></rdfs:Class>
<rdfs:Class rdf:ID=" Computer_Dept">
<rdfs:SubClassOf rdf:resource="#Books"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Electronic_Dept">
<rdfs:SubClassOf rdf:resource="#Books"/>
</rdfs:Class>
…….
<rdfs:Class rdf:ID="Technical_Books"></rdfs:Class>
<rdfs:Class rdf:ID="Java"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="# Technical_Books"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C++"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="# Technical_Books"/>
</rdfs:Class>
<rdfs:Class rdf:ID="DS"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="# Technical_Books"/>
</rdfs:Class>
……
```

The instance of a concept is represented in RDF as follows:

```
<rdfs:Class rdf:ID="CSE"></rdfs:Class>
<rdf:Description rdf:ID="Java">
<rdf:type rdf:resource="#CSE"/>
<Greek Name>OAK</Greek Name>
</rdf:Description>
……
```

4. SEMANTIC INFORMATION RETR I - EVAL SYSTEM

The process of ontology-based information retrieval is as follows:

- Creation of domain ontology by experts;
- Then Collect data from the sources and perform annotation of data with ontology;
- The search engine completes semantic matching of retrieval conditions through ontology reasoning for the user's Search request, then finds out the eligible data set;
- The search result is back to the user.

Here we make use of SPARQL (Simple protocol and RDF query language) for querying against our proposed Semantic retrieval system in figure3, this system searches the books information by the semantic relationships between concepts defined in ontology. The SPARQL query language for RDF provides Semantic Web developers with a powerful tool to extract information from large datasets [11] [12]. The SPARQL has been proposed by the World Wide Web Consortium (W3C) and has recently achieved the recommendation status [13].

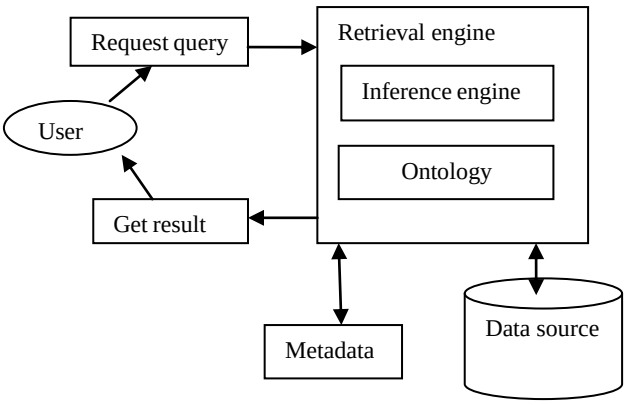

Figure 3: Semantic retrieval system.

4.1 Information Retrieval based on Synonymy Relation

Synonyms are words with the same or similar meanings. Words that are synonyms are said to be synonymous, and the state of being a synonym is called synonymy. The traditional retrieval technology omits the same semantic information, but

if we use the relation of synonyms in ontology, precisions and recall ratio can be improved.
For example, we search for CSE department book information through its name such as "JAVA", using the query statement as follows:

SELECT ?CSE_Dept
FROM Web Data source
WHERE { ?CSE_Dept f:name ?name.
FILTER (?name = "JAVA")}

Using the synonymy relation, the expansion of the statement is as follows:

SELECT ?CSE_Dept
FROM Web Data source
WHERE { ?CSE_Dept f:name ?name.
FILTER (?name = "Java" **OR**
?name = "OAK " **OR**
?name = "Java_Beans")}

4.2 Information Retrieval based on Inheritance Relation

Inheritance relation is the most important semantic relations between concepts. When we search a parent class, all the sub-classes or child classes should be searched out.
For example, when we search information about "Technical Books", the query statement is:

SELECT ?value_of_Technical_Books
FROM Web Data source
WHERE { f: Technical_Books
f: value ?value_of_ Technical_Books}

Using the inheritance relation, the expansion of the statement using is as follows:

SELECT ?value_of_ Technical_Books
FROM Web Data source
WHERE { f: Technical_Books
f: value ?value_of_ Technical_Books **UNION**
f: Java f: value ? value_of_ Technical_Books **UNION**
f: C f: value ? value_of_ Technical_Books **UNION**
f: DS f: value ? value_of_ Technical_Books **UNION**
f: C++ f: value ? value_of_ Technical_Books }

4.3 Information Retrieval based on Part-whole Relation

Part-whole relation is another one of the most semantic relations. When we search information about whole entity, all the parts of information should be searched out. In an ontology, a concept which is defined as aggregation of other concepts is expressed using this relation.
For example, when we search information about "Electronic department books", the query statement is:

SELECT ?Electronic_Dept
FROM Web Data source
WHERE { ?Electronic_Dept f: in f: DSP}

Using the part-whole relation, the expansion of the statement is as follows:

SELECT ?Electronic_Dept
FROM Web Data source
WHERE { ?Electronic_Dept f: in f: DSP **UNION**
? Electronic_Dept f: in f: Lab_Manual **UNION**
? Electronic_Dept f: in f: Theory **UNION**
……}

5. CONCLUSION

With the development of educational system large numbers of books are available today on the web so retrieval of relevant books is a challenging task. Traditional information retrieval system does not provide relevant result according to user needs. This paper has studied semantic retrieval of books information based on ontology and SPARQL. Then we have proposed books ontology model and information retrieval system which will perform intelligent information retrieval through semantic relationship between books concept and this system also gives a mechanism to retrieve synonym words which is the major issue in WWW.

6. REFERENCES

[1] Weihua Li, "Intelligent Information Agent with Ontology on the Semantic Web" 0-7803-7268-9/01 2001 IEEE.

[2] Yuefeng Li and Ning Zhong, "Capturing Evolving patterns for Ontology based Web mining", International Conference on Web Intelligence, 2004.

[3] P. Castells, M. Fernandez, D. Vallet, "An adaptation of the vectorspace model for ontology-based information retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, pp. 261-272, 2007.

[4] J. Li, J. Y. Song, H. Zhong, "Ontology-based query division and reformulation for heterogeneous information integration," Journal of Software, vol. 18, no.10, pp. 2495−2506, 2007 (in Chinese).

[5] J. Arpirez, A. G. Perez, A. Lozano, "(ONTO)2 agent :An Ontology based WWW Broker to Select Ontologies," In: Proc. of the Workshop on Application of Ontologies and Problem-Solving Methods, UK, 1998, pp. 16-24.

[6] Ontobroker. http://ontobroker.aifb.uni-arlsruhe.de.

[7] SKC. http://www-db.stanford.edu/SKC/

[8] M. Karyda, T. Balopoulos , et. Al. "An ontology for secure e-government applications", Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06) 0-7695-2567-9/06 , 2006 IEEE.

[9] Protege: http://protege.stanford.edu/

[10] W3C, Resource description framework (RDF): concepts and abstract syntax, in: G. Klyne, J.J. Carroll, B. McBride (Eds.), W3C Recommendation, 10, February 2004. http://www.w3org/TR/2004/REC-rdf-concepts-200402101

[11] J. Perez, M. Arenas, C. Gutierrez, "Semantics and comp-

lexity of SPARQL," In: Cruz IF, Decker S, eds. Proc. of the ISWC 2006, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 30-43.

[12] R. Angles, C. Gutierrez, "The Expressive Power of SPARQL," In: Sheth AP, Staab S, Dean M, Paolucci M, Maynard D, Finin TW, Thirunarayan K, eds. Proc. of the ISWC 2008, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 114-129.

[13] Prud'hommeaux E, Seaborne A, "SPARQL Query Language for RDF," W3C, 20080115, 2008. http://www.w3.org/TR/rdf-sparqlquery

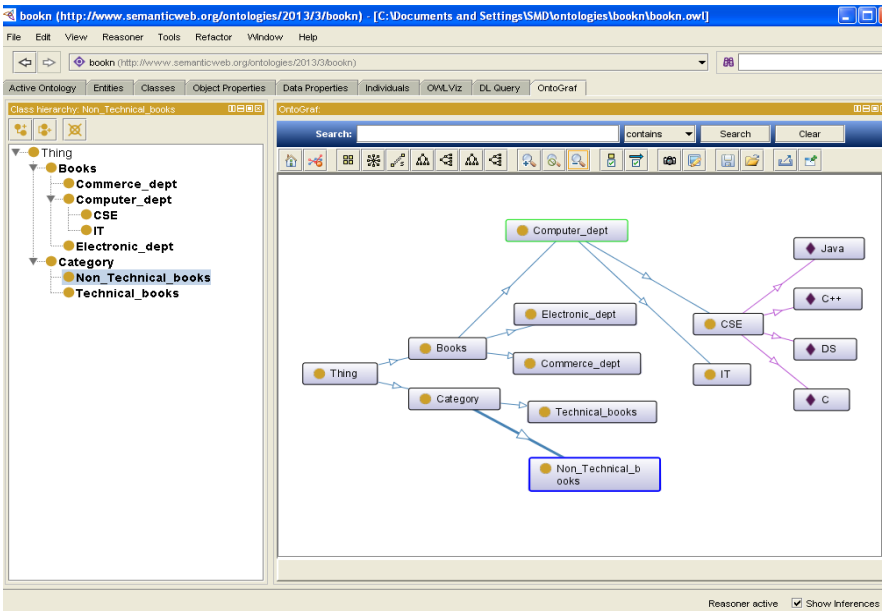

Figure 1: Ontology development using protégé tool.

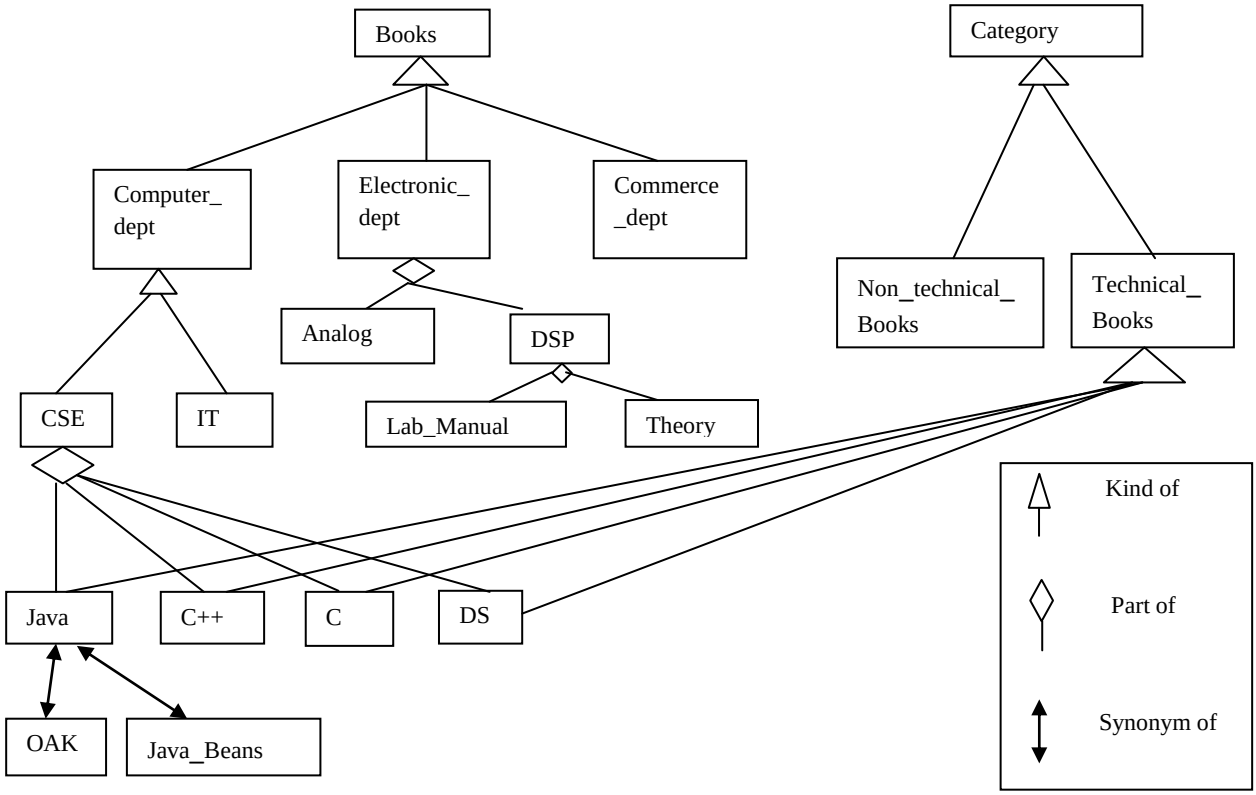

Figure 2: A Small portion of an ontology for books domain.