IJCSC Volume 5 • Number 1 March-Sep 2014 pp.63-67 ISSN-0973-7391

Movie Related Information Retrieval Using Ontology Based Semantic Search

Tarjni Vyas, Hetali Tank, Kinjal Shah
Nirma University, Ahmedabad
tarjni.vyas@nirmauni.ac.in, tank92@gmail.com, shahkinjal92@gmail.com

Abstract— Semantic search has become a grand vision for improving retieval effectiveness in today's scenario. Most of the existing ontology based semantic search models requires user to enter a query in formal languages. It hinders the usability of the retrieval systems. Aiming to solve the above limitations and improve the retrieval effectiveness, a framework for ontology based information retrieval is proposed. In order to overcome the usability limitations, a query interface which requires the user to enter the query in natural language is provided. A domain-specific ontology based on movies is used to develop a prototype of the proposed model which improves search accuracy.

I. INTRODUCTION

Most of todays search engines are keyword-based i.e. search based on literal strings. The main obstacle is that the meaning of present web content cannot be processed by machines. The capabilities of current software to interpret web content and extract useful information are very limited. An alternative approach is to represent web content in a form that is easily processed by machines. This plan to revolutionize the web is semantic web initiative.

A. Semantic Web

The Semantic Web is the extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. It has been described in rather different ways: as a utopic vision, as a web of data, or merely as a natural paradigm shift in our daily use of the Web. Most of all, the Semantic Web has inspired and engaged many people to create innovative semantic technologies and applications. semanticweb.org is the common platform for this community.

B. Ontology

It is a specification of all the relevant concepts and their relationships within a given domain, typically in a hierarchical data structure.

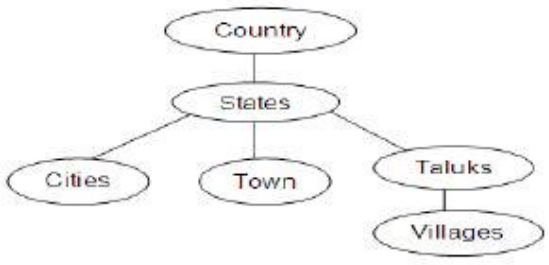

Fig. 1 Ontology for geography of a country

A common set of terms that describes and represents a domain is defined as ontology. It can enhance the functioning of web by improving the accuracy of web searches. An example of th e ontology showing the hierarchical structure of geography of a country is shown in Fig. 1. The search based on ontologies looks for only those pages that refer to a precise concept rather than pages using ambiguous keywords.

C. Ontology

Set of techniques that can be used for retrieving knowledge from structured data sources like ontologies constitutes Semantic Search. When user enters a query using User Interface, the search engine performs a semantic search on KnowledgeBase (KB) (consists of ontologies and RDF files). This semantic search provides the user an organized and much more related data where it uses the synonym/meaning and the search results are displayed. A lot of search time is saved for the users since the actual intended data is presented to them rather than WebPages.

However, one of the most serious problems is that most of the retrieval systems that are based on semantic search do not provide natural language query interface and they want the user to express the query in terms of an ontology based query language. Hence, a natural language query interface that increases the usability of the retrieval system and eases the user to enter the query is essential.

Hence, in this paper, we tackle the problems of syntactic search by providing a framework that accepts natural language queries and retrieves answers using movie ontology. The remainder of the paper is organized as follows. In section II, the related work is being discussed and in section III,

overall architecture of the proposed model is being discussed. Section IV concludes the paper.

II. RELATED WORK

Semantic Search tends to improve retrieval effectiveness. Guha et al [1] designed an application called Semantic Search to improve traditional web searching. Based on the scope of semantic search, it has been applied in different environments. Finin et al [2] discussed about applying semantic search over web so that it improves search effectiveness of information retrieval systems. In Semantic Web area, semantic search system provides search mechanisms over a single KB which is different from standard Information Re-trieval (IR) model that provides document searching. Hence, there is more emphasis on developing new techniques that captures user queries and converts them into formal query representation.

Fernandez et al [5] designed a semantic search system that provides Natural Language Interface (NLI) for users to query and retrieve the results over the semantic web. Linguistic Component as designed by Lopez et al [7] is used to provide NLI. It augments traditional IR approach.

Bernstein et al [3] introduced GINO (Guided Input Natural language Ontology editor) based on this approach. It allows users to edit and query the ontologies in English. It makes use of small grammar rules to provide query suggestions. Once a user enters query, it is converted into Simple Protocol And RDF Query Language (SPARQL) query and is used to search the ontology.

Fernandez et al [8] designed a retrieval system which follows ontology based semantic search approach. The overall retrieval process of the system consists of following steps. The system takes natural language query as input and it is converted into semantic entities by query processing module which has been replaced by cross-ontology ques-tion answering system, PowerAqua. The second step is to retrieve and rank the documents related to users query. For this, documents that are annotated are indexed for retrieval purpose using indexing module which consists of annotation algorithm. The final output of the system is a complementary list of semantically ranked relevant documents and a set of ontology elements that answers users question.

Castells et al [4] designed a retrieval system that exploits ontology based KBs to improve search over large document repositories. Semantic search is combined with traditional keyword based retrieval which tolerates sparseness of KB. The overall retrieval process consists of following steps. This system takes as input RDF Data Query Language (RDQL) query and this is executed against the KB. The output of this step is list of instance tuples that satisfy the query. For this execution, ontology processing library, Jena Toolkit is used. Document Annotation is done using semi-automatic technique. These annotations are given weights based on TFIDF algorithm. The documents that are annotated with the instances returned in previous step are presented to the user. Giunchiglia et al [6] presented an approach called concept search which is search based on computation of semantic relation between concepts. It reuses retrieval model and data structures of syntactic search but the only difference is that words are replaced with concepts and syntactic matching of words is extended to semantic matching of concepts.

The semantic resource used for most of the query answer-ing systems is ontology. One such system called PowerAqua, designed by Lopez et al [7] takes as input natural language query and returns answers retrieved from ontologies found anywhere on semantic web.

III. PROPOSED WORK

A framework for movie ontology based Information Re-trieval model that is expected to improve retrieval effectiveness has been proposed. This framework is depicted in Fig. 2.

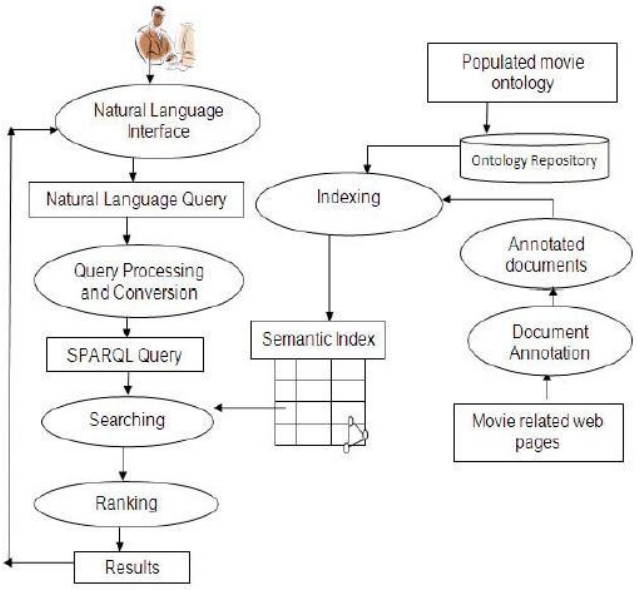

Fig. 2 Framework for ontology based information retrieval

A. Retrieval Process

The overall retrieval process as shown in Fig. 2 involves the following steps.

1) Movie related documents are augmented with metadata

. This process of adding metadata is called annotating a docu-ment. Annotation is done using GATE (General

Architecture for Text Engineering) tool.

2) When a user enters natural language query, query processing is done that converts the query into semantic entities.

3) The SPARQL query is constructed using these semantic entities.

4) This query is used for searching over the triples present in the ontology repository .

5) The indexing process is carried out that result in an inverted index which consists of annotated documents based on entities present in the ontology. This index is stored in a relational database. 6) The inverted index is searched for relevant documents. 7) The documents relevant to the entities are retrieved, ranked and presented to the user. 8) The final output of the system is ranked relevant documents.

B. Movie Ontology

Domain specific (based on movies) ontology is chosen for building a prototype of the retrieval system. It contains classes, subclasses and relationships between them are expressed in terms of properties. This ontology is used as a KB and stored in an ontology repository. One of the popular extensions of ontology files is .owl format. The hierarchical structure consisting of concepts or classes of movie ontology is depicted in Fig. 3.

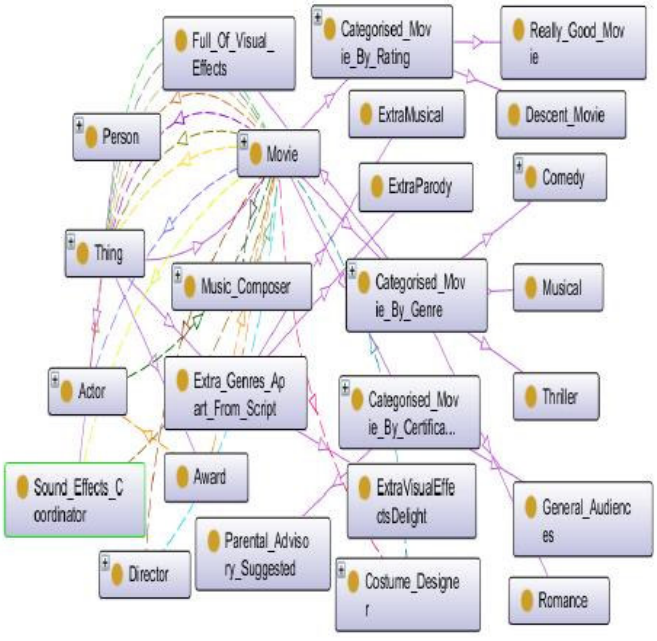

Fig. 3 Hierarchical structure of movie ontology

C. Document Annotation

A semantic annotation in a document is additional information that identifies or defines a concept in a semantic model in order to describe part of that document. It deals with providing metadata to the information contained in a document. Annotations can be manual (performed by one or more people), semiautomatic (based on automatic suggestions), or fully automatic. Manual annotation tools allow users to add annotations to web pages or other resources, and share these with others. Automatic tools can perform sim-ilar annotations (such as named-entity recognition) without manual intervention.

D. Ontology Population

It is a knowledge acquisition technique where instances of ontologically defined concepts and relations are extracted and classified from an information resource. It relies on (semi) automatic methods to transform unstructured (e.g. corpora), semi-structured (e.g.html pages, etc.) and structured data sources (e.g. data bases) into instance data. Movie ontology is populated with instances of actors, directors, etc.For example, the class Movie is populated with following instances:

1) Enthiran
2) Ghajini
3) Thaandavam
4) Alaipayuthe

E. Defining Properties

A property is a directed binary relation that specifies class characteristics. They are attributes of instances and some-times act as data values or link to other instances. Properties may be transitive, symmetric, inverse and functional. They may possess domains and ranges. Two types of properties are:

1) Object properties link individuals to individuals
2) Data type properties link individuals to data values

Some of the object properties defined in movie ontology are:

3) playsin- Rajinikanth playsin Enthiran
4) is directed by- Enthiran is directed by Shankar Thus, RDF triples are formed which consists of subject, predicate and object. For example, Enthiran is directed by Shankar is a triple where Enthiran is the subject, is directed by is the predicate and Shankar is the object. Triple patterns are usually queried using SPARQL query language.

F. SPARQL

It is an RDF query language and a query language for databases that is able to retrieve and manipulate data stored in RDF format. Some of the query forms supported by SPARQL are:

1) SELECT query - used to extract values from a SPARQL endpoint.
2) CONSTRUCT query - used to extract information

from the SPARQL endpoint and transform the results into valid RDF.

3) ASK query - used to provide a simple True/False result for a query on a SPARQL endpoint.

4) DESCRIBE query - used to extract an RDF graph from the SPARQL endpoint. For example, if a user needs to retrieve all the films played by Rajinikanth, then the relevant SPARQL query would be: SELECT? URI WHERE¡http://www.owlontologies. com/Movie.owl/Rajinikanth¿ ¡http://www.owlontologies. com/Movie.owl/playsIn¿ ?URI. Where URI is a variable where results of the query gets stored, Rajinikanth is the subject, playsIn is the predicate. This query retrieves the objects related to the subject and the spredicate present in the query.

G. Functionalities

1) Query Processing and Conversion Query processing operates in domain specific ontology scenario where user terminology is translated into ontology terminology. The users query is converted into semantic triples (subject, predicate and object) format and in turn to SPARQL query. The integration of this conversion part into the framework provides usability since natural language query interface eases the user to enter his query.

2) Indexing In the proposed framework, it is assumed that web pages are indexed using semantic knowledge. Ontology based indexing provides more accurate and precise results. This is achieved by linking unstructured web pages to semantic space by means of explicit annotations. An inverted index is created where the documents in which a semantic entity occurs are stored in a relational database. This database is used for performing semantic search when user enters a query.

3) Searching and Ranking A user interface is provided where user is allowed to enter his/her query. It is converted into SPARQL query. This query then gets posted to server which contains a repository of ontologies. The documents that are relevant to the semantic entities provided in the query are retrieved using the index. Then any of the scoring methods can be used to rank the retrieved documents.

H. Parameters for Evaluation

1) Precision - It refers to the fraction of retrieved documents that are relevant to the search. The formula to find precision is given by,

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

language query interface has been designed so that the user need not have a prior knowledge of ontology based query language which is very complex. This kind of

2) Recall - It refers to the fraction of the documents that are relevant to the query that are successfully retrieved. The formula to find recall is given by,

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

IV. CONCLUSIONS

A solution based on semantic search that addresses the problems of syntactic web is discussed. A comprehensive search model based on ontologies which is used as KB is being designed. This ontology based search model improves search accuracy. A domain-specific ontology (Movie ontol-ogy) has been chosen for experimental purpose. A natural interface tends to improve the usability of the retrieval system.

V. REFERENCES

[1] Guha R.V., McCool R. and Miller E. (2003) Semantic search, Proceed-ings of the 12th International World Wide Web Conference, pp.700-709.

[2] Finin T., Mayfield J., Fink C., Joshi A. and Cost R.S. (2005) Informa-tion retrieval and the Semantic Web, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, pp.414. Belmont, CA: Wadsworth, 1993, pp. 123135.

[3] Bernstein A. and Kaufmann E. (2006) Gino-a guided input natural language ontology, Proceedings of the 5th International Semantic Web Conference ,Athens, GA, USA, pp.144157.

[4] Castells P., Fernandez M. and Vallet D. (2007) An adaptation of vector space model for ontology-based information retrieval, IEEE Transactions on Knowledge and Data Engineering, vol.19, Issue 2, pp.261-272.

[5] Fernandez M., Lopez V., Sabou M., Uren V., Vallet D., Motta E., Castells P.(2008) Semantic search meets the Web, Proceedings of the 2nd IEEE International Conference on Semantic Computing Santa Clara, CA, USA, pp.253260.

[6] Giunchiglia F., Kharkevich U. and Zaihrayeu I. (2009) Concept search, Proceedings of the 6th European Semantic Web Conference Heraklion, Greece, pp.429444.

[7] Lopez V., Sabou M. and Motta E. (2009) Cross-ontology question answering on semantic web-an initial

evaluation, Proceedings of the knowledge Capture Conference, California, CA, USA, pp.17-24.

[8] Fernandez M., Cantador I., Lopez V., Vallet D., Castells P. and Motta E. (2011) Semantically enhanced information retrieval:An ontology based approach ,Journal of Web Semantics,vol.9, pp.434-452.