Multi Domain Information Retrieval using Ontology

Shanmugavadivu. U
Department of Computer Science and Engineering Saveetha Engineering College/ PG Scholar Chennai, Tamil Nadu/India

Mervin.R
Department of Computer Science and Engineering Saveetha Engineering College/ Associate professor Chennai, Tamil Nadu/India

Abstract:- **Web mining is a Knowledge Discovery process from Databases (called KDD) applied to Web data. The amount of content stored and shared on the Web is increasing fast and continuously. Consequently, the ability to access and select relevant information in these huge and heterogeneous masses of data remains a difficult task. Ontologies are special kind of knowledge resources. Ontologies are at the heart of information retrieval from wandering objects, from the Internet. It aim to formalizing domain knowledge in a generic way and provide a common agreed understanding of a domain, which may be used and shared by applications and groups. The system to develop the base framework which will enhance the information retrieval process from the data sources (web) to end user. Multi agent will support the process to work effectively and progress the semantic indexing based on the entity retrieval model. The data object model is adopted for the knowledge base ontology. The customized user interface to get the search keyword and domain. The system will create the search query, parser, rules, validator, mapping and cache memory to retrieve the information and progress the user experience from the knowledge base. The application will be tested in multiple domains like Banking & Finance,Stock &Commodity Market.**

Keywords:*web mining, Ontology, Information Retrievaly.*

I. INTRODUCTION

World Wide Web (WWW) enriches us with enormous amount of widely dispersed interconnected beneficial and dynamic hypertext information.It has furnished the distinct needs of us in various stages like communication, business, entertainment and so on. The current World Wide Web has been reached the peak of its success with respect tovaluable resources of information. The amount of content stored and shared on the Web and other document repositories is increasing fast and continuously. Consequently, the ability to access and select relevant information in these huge and heterogeneous masses of data remains a difficult task. However, most Information retrieval systems have limited abilities to exploit the conceptualizations involved in user needs and content meanings. This involves limitations such as the inability to describe relations between search terms.

In order to overcome these limitations, current Information Retrieval (IR) studies are focusing on relevant documents retrieval using additional knowledge. The main idea is to support a high-level of content and queries conceptual understanding. There are two main categories of conceptual-based information retrieval approaches.

1) The first one concerns approaches that extract semantic meaning from documents and queries by analyzing the latent relationships between text words.
2) The second category consists on approaches that, manually or automatically, construct taxonomy of semantic concepts and relations and map documents and queries onto them.

Ontology, as a knowledge representation, is one of the most used technologies in the second category. The use of ontology in IR is an important parameter to characterizeontology-based methods. The ontology may be used partially through a query expansion . It may also be advanced in both phases of indexing and retrieval. These approaches adopt an advanced use of ontology-based knowledge representation. They can be more efficient especially using domain-information extraction. However, they use specific language for semantic querying which is not easy to be used by the end-users. Formulating a query using such languages requires the knowledge of the domain ontology as well as the syntax of the language.

II. RELATED WORK

Amir Zidi and MouradAbed[1] described generic framework for ontology-based information retrieval. We focus on the recognition of semantic information extracted from data sources and the mapping of this knowledge into ontology. In order to achieve more scalability, we propose an approach for semantic indexing based on entity retrieval model. In addition, we have used ontology of public transportation domain in order tovalidate these proposals.

Keyword-based semantic retrieval system using domain ontology in three phases namely the knowledge phase, the indexing phase and the retrieval phase. We are trying to deal with three main issues of the semantic search and retrieval, Scalability: it involves not only exploiting semantic metadata that are available in data sources but also managing huge amounts of information having a structured and unstructured. In order to achieve more scalability, we propose a semantic indexing approach based on an entity retrieval model. Usability: In order to deal with usability issue, we adopt a keyword-based interface as it provides a comfortable and relaxed way to query about the end-user. Retrieval performance: we are trying to improve the retrieval performance by using a domain-specific information extraction, inference and rules.

Song Jun-feng ,Zhang Wei-ming, Li Guo-hui andXu Zhen-ning[2] described ontology-based information retrieval model for the Semantic Web. By using OWL Lite as standard ontology language, which is a suitable tradeoff between expressivity of knowledge and complexity of reasoning problems, ontology is generated through translating and integrating domain ontologies. The terms defined in ontology are used as metadata to markup the Web's content; these semantic markups are semantic index terms for information retrieval. We can obtain the equivalent classes of semantic index terms by using description logic reasoner. The logical views of documents and user information needs, generated in terms of the equivalent classes of semantic index terms, can represent documents and user information needs.

Domain is a section of the world about which we wish to express some knowledge; domain conceptualization is to abstract a set of terms and a set of knowledge from the domain in terms of the tasks to be solved and the ontological commitment of ontology language used; domain ontology is {the set of domain terms, the set of domain knowledge}, it's explicit specification of domain conceptualization, usually, we use ontology language to write down this specification; ontology is explicit specification of world conceptualization, there is only one ontology about the world, no application needs to use the whole ontology.

In practical application, domain ontology or the integration of several domain ontologies is needed. Suppose the domain D can be divided into n subdomains, then the domain ontology of D can be obtained by integrating domain ontologies of these n subdomains. As different domain ontologies on the Semantic Web are usually encoded in different ontology languages to meet different representation and reasoning needs, we need a translation mechanism that uses OWL Lite as standard ontology language and translate domain ontologies encoded in other ontology languages to domain ontologies endoded in OWL Lite. During the translation, it's inevitable that we will lose some knowledge, that is to say, not all sentences encoded in other ontology languages can be translated into sentences encoded in OWL Lite. Then we integrate these domain ontologies encoded in OWL Lite to obtain domain ontology about all these domains. This domain ontology about all these domains is regarded as ontology, and when new domain ontology is added, it should be firstly translated into domain ontology encoded in OWL Lite, then be integrated with existing ontology, so we have a new ontology which characterizes the world more comprehensively and precisely. The translation mechanism can be achieved semi-automatically with the help of tools such as $OilEd_{P.}$

R.suganyakala and Dr R.R.Rajalaxmi[3] describedMovie related information retrieval using ontology based semantic search . Semantic search has become a grand vision for improving retrieval effectiveness in today's scenario. Most of the existing ontology based semantic search models requires user to enter a query in formal query languages. It hinders the usability of the retrieval system. 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.

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.

Gagandeepsinghnarula and Vishal jain[4] describedImproving statistical multimedia information retrieval(MIR) model by using ontology.The process of retrieval of relevant information from massive collection of documents, either multimedia or text documents is still a cumbersome task. Multimedia documents include various elements of different data types including visible and audible data types (text, images and video documents), structural elements as well as interactive elements. Here have proposed a statistical high level multimedia IR model that is unaware of the shortcomings caused by classical statistical model. It involves use of ontology and different statistical IR approaches for representation of extracted text-image terms or phrases..

III. ARCHITECTURE

Knwoledge Representation is considered as a key feature to represent semantic knowledge. RDF2 schema (RDFS3), which was built upon RDF, was used to develop ontology language. It extends RDF vocabulary with additional classes and properties such as rdfs:Class and rdfs:subClassOf . OWL4 further extends RDFS with additional features such as cardinality constraints, equality and disjoint classes, which enable users to better define their classes. In addition to that, OWL classes may be instantiated by adding new individuals.Entity types defined for RDF, RDFS and OWL.

Indexing is constituted of entities defined for RDF, RDFs and OWL, we designed an indexing system using entity retrieval model. Entity retrieval model A knowledge base, which is constituted of entities defined for RDF, is essentially a labeled and directed graph with the nodes being resources while the edges represent the properties. This graph is essentially a set of RDF Triple (N-Triples). An RDF Triple contains three components each of them is providing complementary pieces of information: subject (node), predicate (property) and object (node). The subject identifies what object the triple is describing, the predicate defines the piece of data in the object we are giving a value to and the object is the actual value.

Once the semantic knowledge is represented and indexed, the next step is querying the EAV graph . In order to do that, we use SIREn6, an efficient semi-structured information retrieval.

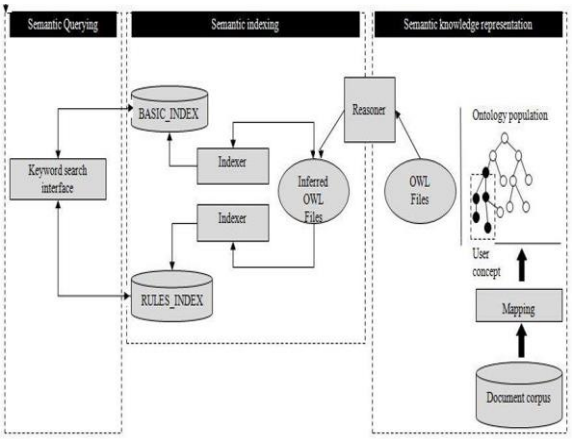


Fig 1: System Architecture

3.1 Ontology for Domains

Ontology is considered as a key feature to represent semantic knowledge. RDF2 schema (RDFS3), which was built upon RDF, was used to develop ontology language. It extends RDF vocabulary with additional classes and properties such as rdfs:Class and rdfs:subClassOf . OWL4 further extends RDFS with additional features such as cardinality constraints, equality and disjoint classes, which enable users to better define their classes. In addition to that, OWL classes may be instantiated by adding new individuals. We can see that user's classes are defined and instantiated based on those entities. It analyzes and extracts data in order to populate the ontology with instances. The next step is inference. The main idea of this step is to expand knowledge base with new added instances using relations and rules . Beyond the relations between classes, used ontology present a set of rules in order to offer better planning to Stocks. As a result, we can have new knowledge about a Stock and commodity Script. After this step, we obtain useful OWL files that will be indexed and used for the search.

3.2 Indexing

As our knowledge base is constituted of entities defined for RDF, RDFs and OWL, we designed an indexing system using entity retrieval model. Entity retrieval model A knowledge base, which is constituted of entities defined for RDF, is essentially a labeled and directed graph with the nodes being resources while the edges represent the properties . This graph is essentially a set of RDF Triple (N-Triples). An RDF Triple contains three components each of them is providing complementary pieces of information: subject (node), predicate (property) and object (node),Thesubjectidentifies what object the triple is describing, thepredicate defines the piece of data in the object we are giving a value to and theobjectis the actual value. In this work, we adopted the Entity Attribute-Value model (EAV model). Before describing our indexing system, we estimated useful to first introduce some basic definitions of EAV model.

3.3 Querying

Three types of queries are supported:

- Full text: keyword-based query when the data structure is unknown. It allows the user to find all the relevant documents that contain all terms in the query using full-text search syntax.
- Structural: when the data structure is known, it produces precise search results using triple patterns to represent partial or complete triples. A triple pattern is a complete or partial representation of a triple <entity, attribute, value>.
- Semi-structural: combination of the two previous query types when the structure is partially known. Full-text search is supported on any part of the triple, which means that the user can use the Keyword-based query syntax to describe his entity, attribute or value.

3.4 Domain search engine using Apache jena API

Jena is an open source Semantic Web framework for Java. Jena has an API to extract data from and write to RDF graphs and OWL ontologies. Model represents a graph in Jena. A model can be created by using data from URLs, files, databases or by combining different sources. In memory and persistent storage for storing large number of RDF triples is provided in Jena. SPARQL can be used to query model. Jena has built in support for many internal reasoners .Pellet reasoner can be used in Jena.

Jena Ontology API Ontologies can be represented by various languages in semantic web ranging from RDFS which is weakest to OWL which is the strongest. Jena ontology API provides a consistent programming interface for ontology application development . Jena ontology API is independent of ontology language used during programming. The Jena Ontology API is language-neutral class names in Java do not mention the underlying language.

OntClass Java class which represent OWL class, RDFS class, or DAML class. Profile is used to establish the differences between the various representations. Every ontology languages are associated with a profile, which contains the details of names of the classes and properties and the permitted constructs. Profile is bounded to an ontology model. OntModel is an extended version of Jena's Model class, which allows access to the statementsin a collection of RDF data. OntModel extends this access by adding support for the kinds of objects in ontology such as classes, properties and individuals.

IV. EXPERIMENTAL RESULT AND ANALYSIS

The performance evaluation is based on the stock market ontology (NIFTY 50). Initially ontology covers 50 stocks which contain the details of the each individual. Information about CEO, Trading info, and result info and experts idea to pick the stock is most used in this experiment.

SPARQL is used to retrieve the information from the ontology. The protégé tool is used to run the generated SPARQL against the stock ontology. The result also displayed in the same tool. As ontology applied the inferred rules

fetching time is reduced. The rules are (A → B) and (B → c) = (A → C).

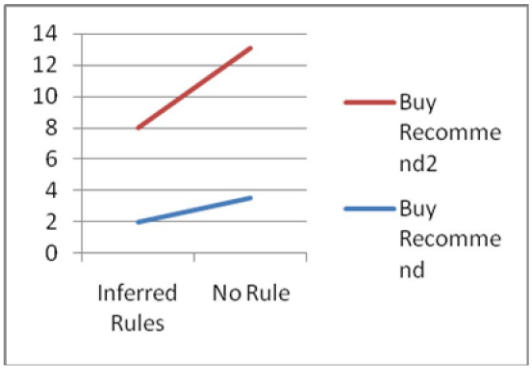
OBJECTIVE

To retrieve all the NIFY 50 stock having the 'BUY' recommendation from the expert. The set of records are extracted from the following conditions: (A is Stock) and (A has an Expert Recommendation) and (A is a NIFTY 50 stock). The SPARQL is generated based on the conditions and applied on the stock ontology.

DATASET DESCRIPTION

Subject	Predicate	Object
Infy	hasCEO	infyCEO
Infy	hasTradingInfo	Infy Trading information
Infy	hasExpertComments	Infy expert comments
InfyCEO	Isa	Person
InfyCEO	Is a	Male
Infy expert comment	Has a	BUY/ SELL
Infy Trading information	Has a	All Trading info

Tab 4.1 Dataset Description



Result Performances Chart

Buy Recommendation result is increased when apply the query on the inferred ontology. Time is calculated by mille seconds. Query is applied in 2 series. Both times getting the improved time and faster result.The average time increment is 166.67% while apply the query on the inferred ontology.

V. CONCLUSION

In this paper, Multi domain framework for information retrieval using ontology and its application in stock market and commodity domain. We tried to exploit the main advantages of semantic knowledge representation by using a domain-specific information extraction, inference and customized rules and also to take advantage of semantic indexing to enhance the retrieval performance. The current implementation can be extended in many ways. Planning to enrich indexed data by using more meaningful rules to better exploit underlying semantics in content being indexed. In addition, we will focus on a new aspect of a personalized search which integrates user's profile in the indexing phase. The main idea is to re-index contents after clustering user's profiles in order to get more relevant matching between well-defined resources and user queries.

REFERENCES

[1] Amir ZIDI and Mourad ABED "A Generalized Framework for Ontology-Based Information Retrieval Application to a public-transportation system"2013 IEEE

[2] Song Jun-feng, Zhang Wei-ming, Xiao Wei-dong, Li Guo- hui, Xu Zhen-ning"Ontology-Based Information Retrieval Model for the Semantic Web "(School of Information System and Management, The National University of Defense Technology, Changsha 410073,China).

[3] R.Suganyakala and Dr.R.R.Rajalaxmi "Movie Related Information Retrieval Using Ontology Based Semantic Search"

[4] Gagandeep Singh and Guru Tegh"Improving Statistical Multimedia Information Retrieval (MIR) Model by using Ontology "Volume 94 – No 2, May 2014

[5] O.Egozi, S. Markovitch, and E. Gabrilovich. 2011. "Concept-Based Information Retrieval Using Explicit Semantic Analysis". ACM Trans.Inf.Syst. 29, 2, Article 8 (April 2011), 34 pages., April 1955.

[6] C. Carpineto and G. Romano. 2012. "A Survey of Automatic Query Expansion in Information Retrieval". ACM Comput. Surv. 44, 1, Article1 (January 2012), 50 pages.

[7] K. Soner K., A. Özgür, S. Orkunt, A. Samet, C. Nihan K., N.A. Ferda.2012. "An ontology-based retrieval system using semantic indexing". Inf. Syst. 37, 4 (June 2012), 294-305.

[8] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, E. Motta.2011. "Semantically enhanced Information Retrieval: An ontology based approach". Web Semant.9, 4 (December 2011), 434-452.

[9] Ayesha Ameen1, Khaleel Ur Rahman Khan2 and B.Padmaja Rani3 "Reasoning in Semantic Web Using Jena " Vol.5, No.4, 2014.