Semantic Web Based Chemical Information Retrieval

A.Kavidha, Dr. A.Saradha

Department of Computer science and Engineering
Institute of Road and Transport Technology
kavitha@irttech.ac.in
Department of Computer Science and Engineering
Institute of Road and Transport Technology
saradha@irttech.ac.in

Abstract— **Information retrieval in chemistry domain is challenging as there are enormous amount of chemical data present in chemical domain. Classification of chemical components will facilitate the scientist to work further. Chemical components are classified based on their structure and properties. Ontologies capture domain knowledge and facilitate easy retrieval. Hence Ontology based information retrieval is applied for chemistry domain. Semantic web technologies are applied in retrieving information from ontologies. In addition to chemical data, drugs details are also encoded in the ontology to find the chemical composition and side effects of drugs. This novel method of retrieval finds application in the domain of medicine.**

Keywords— **Chemical Ontology, Semantic Web, Intelligence Retrieval, Semantic Health Care, Semantic driven IR in Chemistry.**

I. INTRODUCTION

Ontology is Machine-understandable encodings of human domain knowledge and is a formal specification of a shared conceptualization. It provides a common vocabulary of a domain in which the meaning of the terms and the relations between them are defined with different levels of formality. In the computer science domain, ontologies aim at capturing domain knowledge in a generic way and provide a commonly agreed upon understanding of the domain which may be reused and shared across applications and groups. A chemical ontology tries to conceptualize the chemical knowledge in a narrow or broader perspective depending on the granularity level of formalization. In recent years, the importance of building chemical ontology has been felt in the context of evolving Web-based chemistry. However, reports reveal that only a few domain-specific chemical ontologies are developed so far. This may be attributed to the fact that the description of chemical knowledge is difficult because of the fuzziness involved in it, and building a chemical ontology requires considerable effort. Hierarchical organization of data facilitates easy retrieval from knowledge domain.

II. CURRENT SYSTEM AND ITS DRAWBACKS

The World Wide Web has changed the way people communicate with each other and the way business is conducted. It lies at the heart of a revolution that is currently transforming the developed world toward a knowledge economy and, more broadly speaking, to a knowledge society.

At present there is a transition of focus toward the view of computers as entry points to the information highways. Most of today’s Web content is suitable for human consumption. Even Web content that is generated automatically from databases is usually presented without the original structural information found in databases. Typical uses of the Web today involve people’s seeking and making use of information, searching for and getting in touch with other people, reviewing catalogs of online stores and ordering products by filling out forms, and viewing adult material.

These activities are not particularly well supported by software tools. Apart from the existence of links that establish connections between documents, the main valuable, indeed indispensable, tools are search engines. Keyword-based search engines such as Yahoo and Google are the main tools for using today’s Web. It is clear that the Web would not have become the huge success it is, were it not for search engines.

Interestingly, despite improvements in search engine technology, the difficulties remain essentially the same. It seems that the amount of Web content outpaces technological progress. But even if a search is successful, it is the person who must browse selected documents to extract the information he is looking for. That is, there is not much support for retrieving the information, a very time-consuming activity. Therefore, the term information retrieval, used in association with search engines, is somewhat misleading; location finder might be a more appropriate term. Also, results of Web searches are not readily accessible by other software tools; search engines are often isolated applications.

An alternative approach is to represent Web content in a form that is more easily machine-process able and to use intelligent techniques to take advantage of these representations. We refer to this plan of revolutionizing the Web as the Semantic Web initiative. It is important to understand that the Semantic Web will not be a new global information highway parallel to the existing World Wide Web; instead it will gradually evolve out of the existing Web.

Drawbacks of Existing System

- ❖ High recall, low precision. Even if the main relevant pages are retrieved, they are of little use if another 28,758 mildly relevant or irrelevant documents are also retrieved. Too much can easily become as bad as too little [10].
- ❖ Low or no recall. Often it happens that we don't get any relevant answer for our request, or that important and relevant pages are not retrieved. Although low recall is a less frequent problem with current search engines, it does occur.
- ❖ Results are highly sensitive to vocabulary. Often our initial keywords do not get the results we want; in these cases the relevant documents use different terminology from the original query. This is unsatisfactory because semantically similar queries should return similar results.
- ❖ Results are single Web pages. If we need information that is spread over various documents, we must initiate several queries to collect the relevant documents, and then we must manually extract the partial information and put it together

III. SEMANTIC WEB TECHNOLOGIES

The Semantic Web is a "web of data" that enables machines to understand the semantics, or meaning, of information on the World Wide Web.[1] It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. The term was coined by Tim Berners-Lee,[2] the inventor of the World Wide Web and director of the World Wide Web Consortium, which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processe directly and indirectly by machines."

The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it.[3] These technologies include the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

Many of the technologies proposed by the W3C already exist and are used in various contexts, particularly those dealing with information that encompasses a limited and defined domain, and ***where sharing data is a common necessity***, such as scientific research or data exchange among businesses. In addition, other technologies with similar goals have emerged, such as micro formats. However, the Semantic Web as originally envisioned, a system that enables machines to understand and respond to complex human requests based on their meaning, has remained largely unrealized and its critics have questioned its feasibility.

In the context of the Web, ontologies provide a shared understanding of a domain. Such a shared understanding is necessary to overcome differences in terminology. One application's zip code may be the same as another application's area code. Another problem is that two applications may use the same term with different meanings. In university A, a course may refer to a degree (like computer science); while in university B it may mean a single subject (CS 101). Such differences can be overcome by mapping the particular terminology to a shared ontology or by defining direct mappings between the ontologies. In either case, it is easy to see that ontologies support semantic interoperability. Chemical classes, the objects found within a chemical classification system, group together chemical entities in a meaningful, scientifically relevant hierarchy [5]. Web searches can exploit generalization/specialization information. If a query fails to find any relevant documents, the search engine may suggest to the user a more general query. It is even conceivable for the engine to run such queries proactively to reduce the reaction time in case the user adopts a suggestion. Or if too many answers are retrieved, the search engine may suggest to the user some specializations. OWL is a richer vocabulary description language for describing properties and classes, such as relations between classes (e.g., disjointness), cardinality (e.g., "exactly one"), equality, richer typing of properties, characteristics of properties (e.g., symmetry), and enumerated classes [11].
At present, the most important ontology languages for the web are

- ❖ RDF is a data model for objects ("resources") and relations between them. It provides a simple semantics for this model; and these data models can be represented in an XML syntax
- ❖ RDF scheme is a vocabulary description language for describing properties and classes of RDF resources, with a semantics for generalization hierarchies of such properties and classes
- ❖ OWL is a richer vocabulary description language for describing properties and classes, such as relationships between classes.
- ❖

IV. ONTOLOGY BASED FRAMEWORK DESIGN FOR CHEMISTRY INFORMATION RETRIVAL

This paper describes a semantic web application that uses an ontology which is created using OWL (Web Ontology Language) with protégé tool[6]. The web application has several modules such as getting queries from user in natural language format, a SPARQL query generator for the natural language query, a module for posting the queries to semantic agent along with ontology and getting results, etc., as in Fig-1.The user enters simple natural language queries in the text box provided as shown in Fig-3. Keywords sets are compared to identify the actual query which is then converted to machine understandable SPARQL. Knowledge base is referred by the agent for the requirement present in the SPARQL. Finally, the result set produced by the agent and then the result is formulated as per the user required output format.

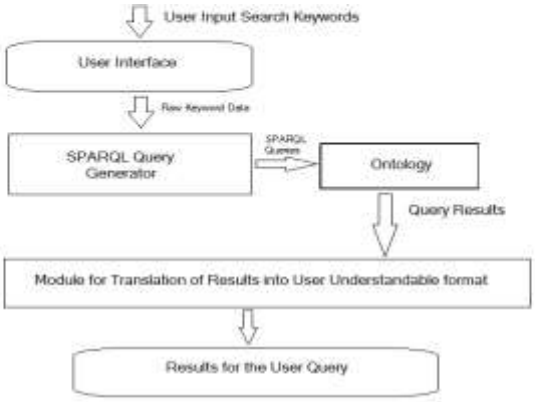

User Input Search Keywords
User Interface
Raw Keyword Data
SPARQL Query Generator
SPARQL Queries
Ontology
Query Results
Module for Translation of Results into User Understandable format
Results for the User Query

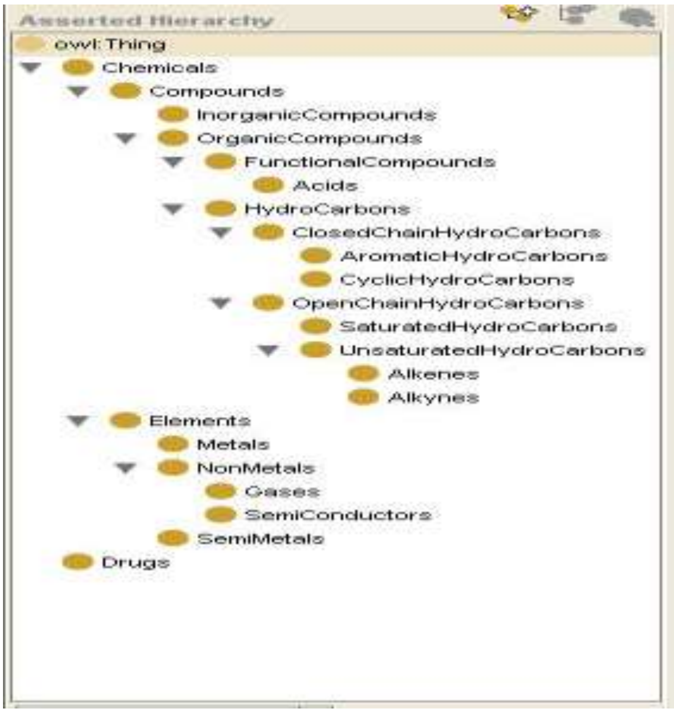
Fig-1 Framework for Semantic Chemistry IR

A.Development of Chemical Ontology

- ❖ The ontology is developed using the protégé version protégé 3.4.2 for this paper[6]. The Steps involved in development of the Chemical ontology are
- ❖ Identification of Classes for ontology
- ❖ Deriving the class hierarchy
- ❖ Identification of the data and object properties of each class
- ❖ Defining data types (domain and Range specification) for properties
- ❖ Creating individuals for each classes
- ❖ Adding values for properties of each individuals
- ❖ Checking the ontology for consistency
- ❖ Testing the ontology by Providing SPARQL
- ❖ Verifying weather it gives the correct results

B. Deriving the class hierarchy

A class can be thought of as a set of similar elements or objects. The relationships typically included in hierarchies of classes. A hierarchy specifies a class C to be a subclass of another class D, if every object in C is also included in D. By default owl: thing is the root class of the entire user defined classes and their subclasses. The Chemical ontology developed as class hierarchy is shown in Fig-2. Compounds, Elements and Drugs are the super classes under owl: Thing root class. Organic compounds are explored a lot compared to In-Organic compounds.



Asserted Hierarchy

- owl:Thing
  - Chemicals
    - Compounds
      - InorganicCompounds
      - OrganicCompounds
        - FunctionalCompounds
          - Acids
        - HydroCarbons
          - ClosedChainHydroCarbons
            - AromaticHydroCarbons
            - CyclicHydroCarbons
          - OpenChainHydroCarbons
            - SaturatedHydroCarbons
            - UnsaturatedHydroCarbons
              - Alkenes
              - Alkynes
    - Elements
      - Metals
      - NonMetals
        - Gases
        - SemiConductors
      - SemiMetals
    - Drugs

Fige-2 Ontology for Chemistry

C. Deriving the data and object properties for each class

Once we have classes it is important to impose restrictions for them. The RDF statements in the ontology can be viewed as a three valued statement i.e subject, predicate and object. Here subject is imposed as domain, predicate is viewed as property and object is viewed as range. Properties are of two types

- ❖ Data property
- ❖ Object Property

Data property

Data type property links an individual with a set of values. i.e., here domain is an individual of any class and range is a set of values of a particular data type. The chemical ontology includes the following data properties.

- ❖ AtomicNumber
- ❖ AvailableForms
- ❖ Banned
- ❖ BoilingPoint
- ❖ Ingredients
- ❖ IUPACName
- ❖ MassNumber
- ❖ MeltingPoint
- ❖ Name
- ❖ Phase

- SideEffects
- Symbol
- Toxicity
- Uses

Object Property

Object property relates an individual with another individual i.e., here both the Domain and Range of the three valued statement are individuals.

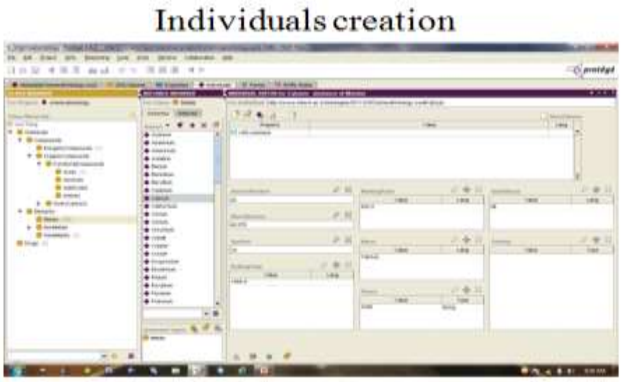
The list of object properties included in ontology is given below.

- ParentOf
- DerivedFrom

The domain here is a chemical and the range is another chemical.

D. Creating individuals for each class

Individual objects that belong to a class are referred to as instances of the class and is shown in Fig-3. In this step, individuals for each class are created and their respective data property and object property values are filled. This is actually called as the knowledge acquisition phase.



Individuals creation

Fig-3 Individual Creation

E. Checking the ontology for consistency

The developed ontology should be checked for consistency and it should be classified for the purpose of error detection. Classifying the ontology is must for querying purposes. The Pallet reasoner plugin of protégé is used for reasoning the ontology and to check for inconsistencies.

F. SPARQL Query

Protégé provides SPARQL Query tab for posting queries. It has been used at the development time to get various inferences and to test feasibility of IR. Later, user can use portal for giving the natural language query which then translated into SPARQL.

G. Implementation of Chemistry IR

The Framework consists of three phases namely,

- Query Input
- SPARQL Query generation
- Query Result Display

The IR System is shown in Fig-4 below:-

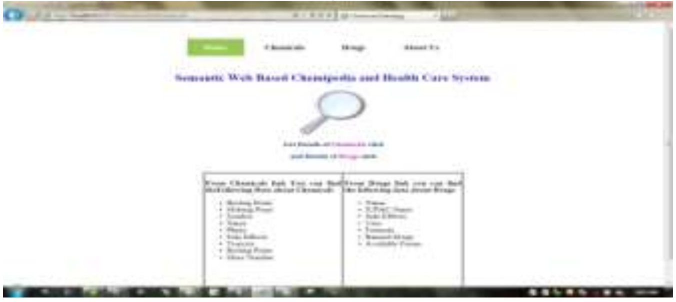
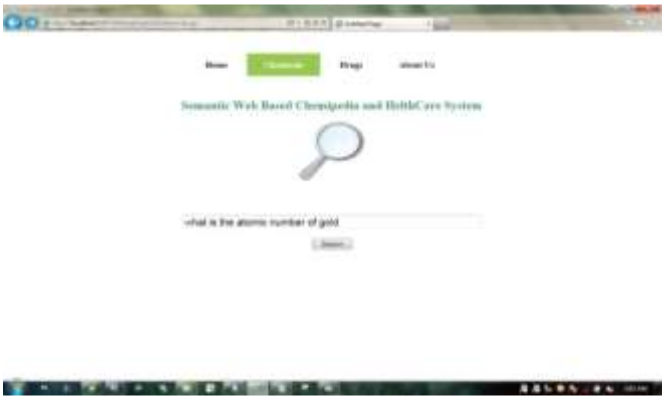


Fig-4 Semantic web based Chemical information retrieval system

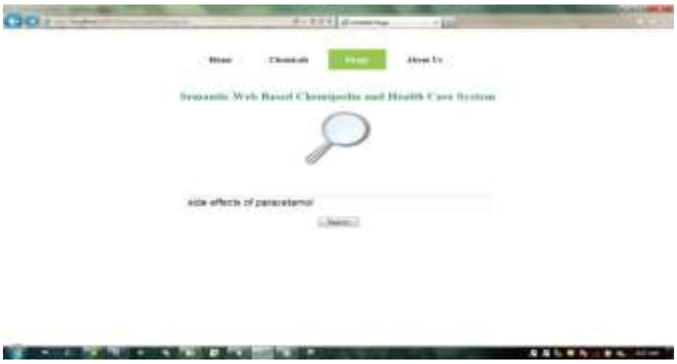
H. Query Input

The Query input module for chemical IR Framework allows the user to get the natural language query. The UI of the system is in Fig-5 and the UI for Drug Information system is shown in Fig-6. The query is then passed to the semantic query generator module.



what is the atomic number of gold

Fig 5 Semantic Chemistry IR-Query in Natural language



side effects of paracetamol

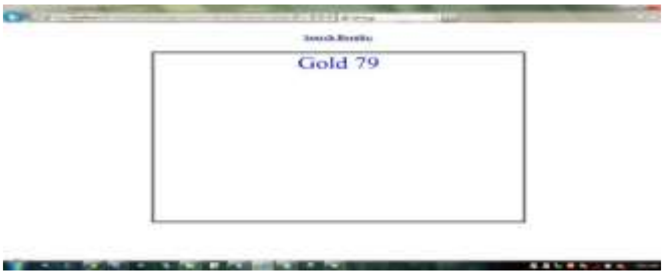
Fig-6 Health Care system

I.SPARQL Query Generator

The SPARQL query generator translates the natural language query into equivalent SPARQL query. The generator is capable of processing only a limited number of word sequences. The query posted by the user is translated by the SPARQL query generator and then the Semantic Agent uses the ontology to give the inference the user. The result is then formulated as per requirement. Open source semantic web framework for building Semantic WEB called JENA is used for Agent implementation.

K. Query Result Display

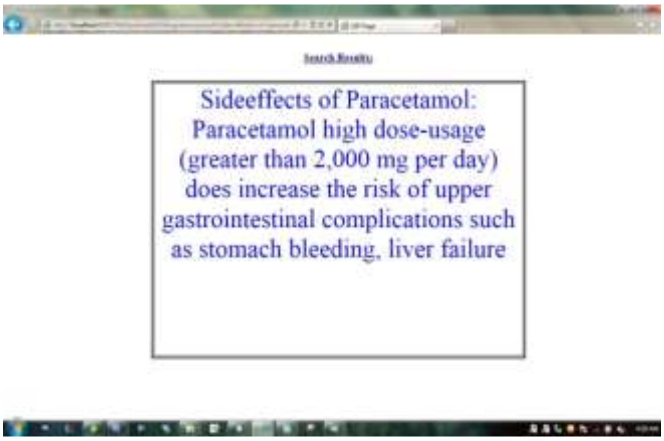
SPARQL query sent to the agent is parsed and the program understands itself the relationship between triples in the knowledge base. The ontology gives the guidance necessary to relate the objects which proves the inference produced. The result obtained for the query in the above Figure-5 is shown in Figure-7 and the result of query in Figure -6 is shown if Figure-8.



Search Results

Gold 79

Fig-7 Result of the Query



Search Results

Sideeffects of Paracetamol:
Paracetamol high dose-usage
(greater than 2,000 mg per day)
does increase the risk of upper
gastrointestinal complications such
as stomach bleeding, liver failure

Fig-8 Inference for Query in health care system

V.CONCLUSION

Ontology-based systems are that they make use of the semantic information to interpret and provide precise answers to questions posed in NL and are able to cope with ambiguities in a way that makes the system highly portable [3]. In this paper, the details of Chemicals and drugs are maintained as ontology and mined for the results based on the user's request. The semantic agent automatically searches for the relationships as per the query components with the consultation of knowledge base. Any number of new complex queries may be formed by the system as opposed to the database technology. The results returned can be justified with the help of inference rules. Administrator can update the ontology without reconstructing the whole system.

REFERENCES

1. Mariano Fernindez Lopez, Asuncion Gomez-Perez, and Juan Pazos Sierra. "Building a Chemical Ontology Using Methontology and the Ontology Design Environment", Polytechnic University of Madrid Alejandro Pazos Sierra, University of Coruiia , IEEE INTELLIGENT SYSTEMS(1999)
2. Punnaivanam Sankar, and Gnanasekaran Aghila "Design and Development of Chemical Ontologies for Reaction Representation", J. Chem. Inf. Model., Vol. 46, No. 6, 2006
3. Vaness , Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. "Is Question Answering fit for the Semantic Web?: A survey, Semantic Web 2" (2011) 125–155, DOI 10.3233/SW-2011-0041,IOS Press
4. Susie Stephens, *Oracle* Alfredo Morales and Matthew Quinlan, *Cerebra, "*Applying Semantic Web Technologies to Drug Safety Determination",. IEEE Intelligence System (2003)
5. Hastings et al. ," Structure-based classification and ontology in Chemistry", Journal of Cheminformatics, 2012
6. *http://www. protege.stanford.edu*
7. *http://en.wikipedia.org/wiki/chemicals*
8. *http://lentech.com/periodic/elements*
9. *http://www.drugbank.ca/drugs/*
10. "Melting and Boiling Point Tables,Vol-I",Carnellay,
    a. Thomas, University of Toronto
11. Grigoris Antoniou and Frank van Harmelen, "A Semantic Web Primer", 2nd Edition
12. Morgan Kaufmann – "Semantic Web for the working Oncologist Effective Modelling in RDFs and OWL" (2008)