Domain Specific Semantic Web Search Engine

KONIDENA KRUPA MANI BALA[1], MADDUKURI SUSMITHA[2,] GARRE SOWMYA[3], GARIKIPATI SIRISHA[4], PUPPALA POTHU RAJU[5]
[1,2,3,4]*B.Tech, Computer Science, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India*

Abstract -- The World Wide Web is the repository for huge data which comprise text documents and other web resources. These documents are identified by Uniform Resource Locators (URL) and are accessed via internet. The volume of information grows in billions every second. So retrieving accurate information for users query is difficult. Search engines use keyword based search methodology to retrieve information, which may fail to provide accurate information. Semantic web technologies play important role to overcome this problem. In this paper we propose an ontology based domain specific semantic web search engine. The search engine hosts information about agriculture domain.

Index Terms: Semantic Web, Web Ontology, Information retrieval, Resource Description Framework (RDF)

I. INTRODUCTION

A Search engine is a platform used to search the information in the World Wide Web. The information retrieved may be in the form of document, images etc. Any search engine uses a keyword based search methodology where each word from the query provided by the user is taken and compared with the matching words in the database. Any document which matches the keyword is retrieved irrespective of the searchers intent. This creates ambiguity in the information retrieved and the user has to go through multiple links. The links are also provided based on the ranking of the document. If the user spends more time reading the link, search engine considers it as relevant link and ranks the link. Traditional search engine does not provide reliability to the users query. For example when the user searches for the information about the suitable crop for a particular season, although the search engine provides thousands of result there is no reliability of the information. The user has to go through multiple links to find relevant information. Another factor is the accuracy of the information is not up to the mark. In order to increase the efficiency in the information retrieved we propose ontology based semantic web search engine. By using semantics we can retrieve relevant information for the user query.

II. SEMANTIC SEARCH ENGINE

For most of the internet users search engine has become the starting point for searching the information in the web. Majority of the users follows keyword based search for information retrieval. In keyword based search, search engine searches its enormous database for the keyword. Indexing is used to organize the database. Then it returns relevant pages for the given query. But by using this type of search users retrieve irrelevant web pages for the given query.

Information retrieval using search engine is not a fresh idea it has different challenges when compared to general information retrieval. Every search engine like Google, yahoo, Bing returns different results due to indexing process. But some researchers started searching information by analyzing the semantics of the given statement. Till now, no search engine could produce accurate results for the given query. Due to the lack of semantic structure present search engines does not provide accurate results. It is difficult for a machine to understand the query and perform indexing. Many researchers have a problem like how a search engine can map the query to the documents where information is available. This can be solved by semantic annotations where we can retrieve meaningful information. To understand the architecture a different architecture is provided.

The first layer URI and Unicode follows important features of existing WWW.URI is a string of standardized form which is used to identify different documents.XML is extensible mark-up language which is used for the documents having structured format. RDF is a framework for representing the information about the documents in graph format. To allow standard description, a RDF Schema (RDFS) was created together with its semantics within RDF.

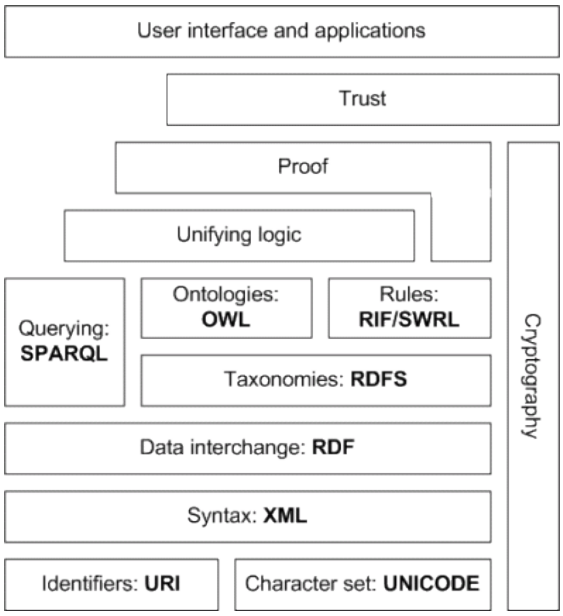

Figure 1: Semantic Web Architecture

RDFS also uses different techniques for describing classes and properties and also used to create light weight ontologies. It also has the capability to extend the definitions for some of the RDF elements. OWL is a language derived from description logics, and offers more constructs over RDFS. For querying RDF data, SPARQL is available. SPARQL is similar to SQL language but it uses RDF triples for matching the query and returning the results. All the semantics are executed in the layers before proof. Formal proof with trusted inputs for the proof results that the outcomes for the query can be trusted. For reliable inputs cryptography is used such as digital signature for the verification of sources.

III. EXIXTING SEMANTIC SEARCH ENGINES AND LIMITATIONS

Hakia

Hakia is a semantic search engine that uses concept based search rather than keyword word match or popularity ranking to bring relevant results. The results are retrieved based on sentiment match rather than popularity of search terms. Another major aspect of Hakia is that it gets information based on matching equivalent terms also, for example "farming=cultivation". Information can be of any form like Web, News and Blogs which can be re-listed according to relevance and date. Hakia uses three evolving technologies. OntoSem is a repository where words are categorized into various senses they convey which form a linguistic database. QDEX (Query indexing technique) it retrieves all the queries relating to the content and a semantic ranking algorithm is used to rank the content. Limitation of Hakia is that it searches for the related content from galleries and videos which increases the search time.

DuckDuckGo

The main feature of DuckDuckGo which no other search engine has is that, if a term has more than one meaning. The search engine will give an option to choose between different meanings. For example if the user searches for Apple it will provide an option to choose among the possible meanings i.e. fruit, Software Company, bank.

Powerset

Powerset is the Microsoft acquired search engine. It uses natural language processing to understand the searches intent and retrieves the pages containing relevant information. It is ultimate search engine to search for the content in the Wikipedia since all the search results are from Wikipedia. Powerset generalizes the information from different sources. Since the information is retrieved from Wikipedia the results are also limited to Wikipedia.

Kngine

It is a semantic search engine which results either web results or image results. It results all the links related to user query. If the user enters the query related to films then Knigne returns information about films, film trailer, review, quotes. If the user searches about the city then it returns local attractions in the city, events, Weather in the city, Hotels.

IV. ONTOLOGY

Ontology can be defined as explicit specification of a concept. Here explicit specification of conceptualization means that, ontology is description of the concepts and the relationships that can exist between the concepts. Therefore ontology provides a vocabulary for representing and communicating knowledge about some topic and set of relationships that hold among the terms in the vocabulary [2]. Here

in this paper we propose to develop an ontology for agriculture related concepts by developing a hierarchy between the concepts and their inter relation.

Why develop ontology?

- To retrieve meaningful information for users query.
- To provide machine understand ability.
- To provide interoperability.

In our project we used RDF which means Resource Description Framework to represent knowledge. In order to represent a concept in terms of its subclasses and relations we define Ontology.

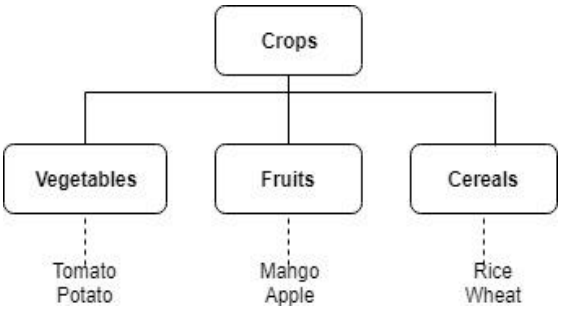

Figure 2: The crops ontology: Crops is the superclass and the subclasses are Vegetables, Fruits and Cereals

As the domain in our project is agricultural domain, the above figure depicts the hierarchy where crops is the main class and vegetables, fruits and cereals are the subclasses of crops.

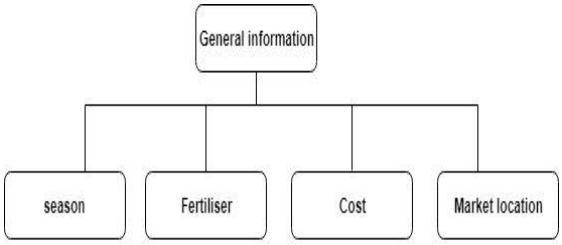

Figure 3: The general information ontology: General information is the superclass of the subclasses season, fertiliser, cost,market location.

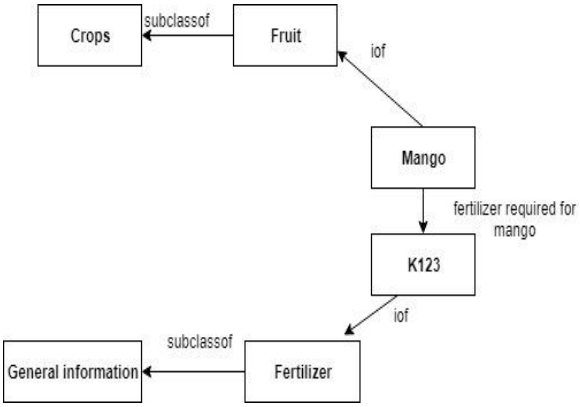

Figure 4: Some classes, instances and relations among them in the crops domain : crops, fruit ,fertiliser, general information are the classes and iof indicates instance of.

V. PROPOSED SYSTEM

We first developed RDF pages based on the agriculture ontology. Resource Description Framework is used to interchange data on the web. It is mainly based on making the statements about resources in triple format i.e. subject-predicate-object format. In triple format subject indicates resource, predicate indicates traits and maintains relationship between subject and object.

An Example to understand RDF format is Alice is in New York. In the above example Alice is the subject, is in is the predicate, New York is the object. RDF statements are represented using graph format.

Resource: The main subject which the RDF expression represents is called the resource. Each resource is associated with a unique id called the URI (Uniform Resource Identifier). A resource may be a part of the web page or an entire web page.

Property: Property is the aspect or a characteristic. Sometimes a property may be resource since it has its own properties. Property mainly describes the attributes of a resource.

An RDF graph consists of these triplets. An example of RDF graph is shown below.

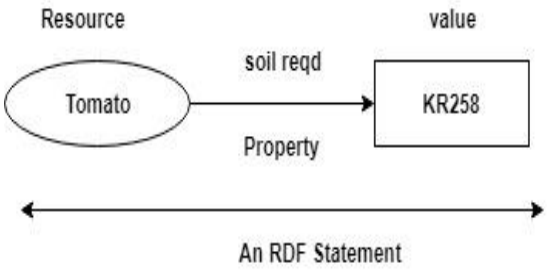

Figure 5: An RDF Statement

```
<?xml version="1.0"?>

<vegetable rdf:ID="Tomato"

xmlns:rdf="http://www.w3.org/199

9/02/22-rdf-syntax-ns#"

xmlns="http://www.india.org

/crops#">

<soilreq>KR258</soilreq>

</vegetable>
```

The above code describes the RDF statement for the resource "Tomato" which is the instance of the subclass Vegetable described in the figure 2. KR258 is the value of the property "soilreq".

RDF SCHEMA: RDF Schema is used to describe groups of related resources and the relationships among the resources [2]. RDF schema is used to define classes and their relationships and also to define properties and associate them with classes.

In RDF Schema each resource can be identified with RDF URI reference which can be described by properties. Generally prefix "rdfs:" is used to indicate the term is RDF Schema. The root class in RDF Schema is "rdfs:resource". "rdf:type" is an instance of rdf:Property (class of RDF properties), and it means that a resource is an instance of a class[2]. rdfs:subClassOf is an instance of rdf:Property which can be used to represent the property of a subclass.

VI. DESIGN OF SEMANTIC WEB SEARCH ENGINE

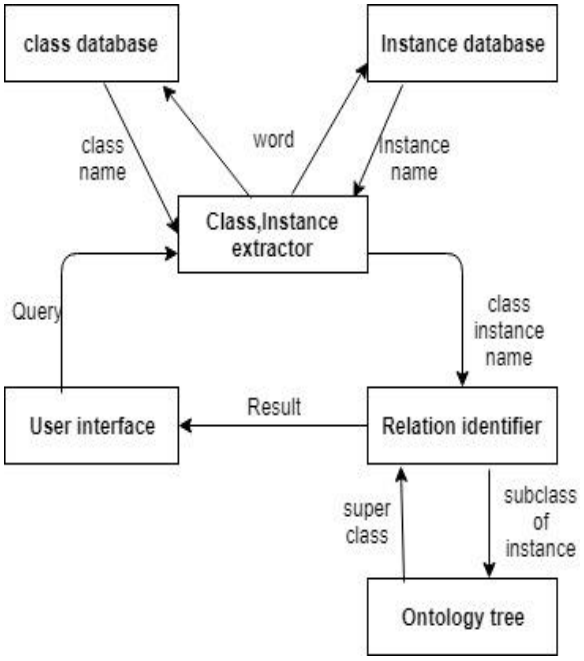

Figure 6: Basic components of semantic search engine

There are five main parts in the design of semantic web search engine

1) A class, instance extractor
2) A class database
3) An instance database
4) An ontology tree
5) A relation identifier

The class instance extractor: It takes input as user query and returns output as class or instance name. The name itself represents that this module is responsible for the extraction of class or instance names. For example if the user query is season required for mango then it extracts the words season and mango. It is done with the help of class and instance database.

Class database: It contains all the class names and their synonyms used in RDF code. It is represented using two columns. One is for class name and other is for synonym. It checks each word given by class instance extractor. If the word is matched it returns yes, and returns the word to class instance extractor

Instance database: It is similar to class database except it stores only Instance names.

Ontology tree: It is used to find out class of an instance or class of a subclass.

Relation identifier: This module is used to find the relationships between class and instances

VII. CONCLUSION

Even though multiple search engines are available to search for information in WWW, the amount of data available in WWW grows numerously. Providing relevant search information to the web searcher has become difficult. Hence search engines which are based on the semantic web technology can effectively handle the problem. The proposed system is based on agriculture domain. This search engine helps to reduce the amount relevant information retrieved. It reduces the need for the user to go through multiple links in order to find appropriate information. The time required by semantic search engine is less when compared to normal search engine.

This search engine is based on the agriculture domain and it is scalable. It can also be used to other domains and only requires to feed the relevant RDF codes of the particular domain.

Finally semantic search engine can reduce the ambiguity for the information retrieved for the users query, the users can avoid going through multiple links.

REFERENCES

[1] Berners-Lee.T,1999. Weaving the Web:The Original Design and Ultimate Destiny of the World Wide Web by its Inventor,New York:Harper SanFrancisco.

[2] A Domain Specifi Ontology Based Semantic Web Search Enginehttps://arxiv.org/ftp/arxiv/papers/1102/1102.0695.pdf

[3] W3C, Resource Description Framework (RDF) Model and Syntax Specification, W3Crecommendation February 1999, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

[4] "Google Search Engine".http://www.google.com

[5] David E. Goldschmidt and Mukkai Krishna Moorthy; Architecting a Search Engine for the Semantic Web