

Prototype for Automated Ontology Construction for Semantic Search in Health Care Management using Herbal Plants

Kripanshu Bhargava

B. Tech Student

*Department of Information Technology
SRM University, Kattankulathur,603203, India*

Debashri Mukherjee

B. Tech Student

*Department of Information Technology
SRM University, Kattankulathur,603203, India*

Vadivu. G

Professor

*Department of Information Technology
SRM University, Kattankulathur,603203, India*

Abstract

The Semantic Web is deemed to be the next major step in associating information in the web. It facilitates the linkage of data between two related sources and to be understood by the machine. Ontology is the essential building block of the semantic web for making the machine to communicate with other machines. Constructing the ontology for the general domain is very tedious and the construction of Medicinal plant ontology to be one of the most useful ontology for the average person to understand the usage of medicinal plants. Earlier research on ontology design methodologies shows that constructing ontology using manual process is tedious, but it is difficult to create the ontology automatically, since the sources of the input are in different forms, like, structured, semi structured, or unstructured data. In this paper, we present an automated ontology construction methodology for the medicinal plants. Java is used to extract the Wikipedia web pages related to medicinal plants, Java enabled application (Jena) is used for the automated construction of ontology, and Protégé is used for verifying the consistency of the ontology. Jena with SPARQL is used for inferring the usage of Indian medicinal plants. An Automated Knowledge framework for the Indian Medicinal Plants (AKIMP) is developed based on this proposed concept. The system is exhibited for fifty thousand concepts and gives the related information from Traditional Knowledge Digital Library (TKDL)[12] and Wikipedia web pages. Thus, it is possible to construct the automated knowledge base to retrieve the required resultant webpage based on the users' need. This automated ontology construction methodology is an improvised method in analyzing the use of medicinal plants.

Keywords: Semantic Web, Resource Description Framework (RDF), Web Ontology Language (OWL), Jena, SPARQL Protocol and RDF Query Language (SPARQL)

I. INTRODUCTION

The development Automated Knowledge framework for the Indian Medicinal Plants (AKIMP) of the prototype includes the formation of the ontology and a simplistic interface for querying the system. Jena is used to build semantic web applications with the ontology representation of Resource Description Framework (RDF) and Web Ontology Language (OWL). Various query patterns are fetched using RDF Query Language and SPARQL Protocol. Lexical terms are considered to enable automated construction.

Throughout Indian folklore and mythology, medicinal plants have been shown to cure a large slew of diseases. Even though, the effects of these assertions are still to be proven by modern science, their effects have a large set of believers. In the modern world it has been realized that the herbal drugs strengthens the body system without side effects.

Web is having large volume information related to herbal plants and becomes very time consuming process to search for user specific information. Searching for the specific information by the average user is a difficult process. Search engines are used to search for these documents, but they still have to be processed by themselves before any useful information can be extracted. As text based information, there are many limitations in using the medicinal plants search:

- Searching text-based documents is a time consuming process. It provides general information, which is not always required by the user.
- There isn't any method to construct the ontology automatically. This study is used to address these limitations to providing useful information.

Semantic Web in World Wide Web is to collect, manipulate and annotate the information by providing categorization, uniform access to resources and structuring the information in machine process-able format. To organize the information in machine understandable form, Semantic Web has introduced the concept of "Ontology" (Antoniou and Harmelen, 2004).

During the past years, researchers and developers worldwide have done a wide range of ontology construction works. Those ontologies were created manually/semi-automated and there is no model to construct fully automatically. Ontologies are used to compile and construct knowledge in research areas and adapt machine-comprehensible semantics. Currently, a consolidated method for ontology construction does not exist, and in order to formally construct ontologies in a reusable manner, an automatic ontology construction method is required[3]. Automated ontology construction requires the input from domain experts, and it is not possible for the researchers to have the knowledge in specific domain. This paper focuses on how to design and implement a prototype system for automated ontology building.

Documentation of this existing knowledge, available in public domain, on various traditional systems of medicine has become imperative to safeguard the power of this traditional knowledge.

Ontology explains the concepts, their relationships and properties in their respective domain and it can be used both to offer electronic inference. This is a relevant perception for management of knowledge. Ontology grants understanding of the information organization. With help of a common ontology, information that is distributed in many different applications and documents can be seen in an easy way to understand and navigate. It becomes very easy to search implicit and explicit information with the help of ontology, thus it helps in bridging the gap between the implicit and explicit knowledge. The advantages of ontology are: sharing of knowledge, logic inference and rephrasing the knowledge.

Ontology defines a common phraseology for researchers who want to share information in a domain. It includes machine-interpretable explanation of basic concepts in the domain and connections among them. In the pragmatic aspect, developing ontology includes: classification in ontology, arranging the classes in a taxonomic hierarchy, defining properties, and loading in the values for properties of instances.

A. Related Works

In this paper[1], they introduced the method of automatic ontology construction research, and several kinds of technology on ontology construction. By analysis, the conclusion is obtained as shown in the following: though there are many ontological construction methods and editing tools, most of them are only instructional, but are not validated. Then, they constructed the ontology in military intelligence by using the thesaurus and data base resources. Finally they achieved the automatic construction of domain ontology. The constructed ontology after manual modification can achieve the purpose of application. The ontology has good scalability, and can be further enriched and could be improved.

In this research[2], a support system for Vietnamese ontology construction was proposed. The methodology was derived from a combination of methods based on statistics, lexical patterns and patterns of frequent sequences. The shortcomings of each individual method can be overcome by the unified approach. This can help avoid missing relations in the detection task. A real Vietnamese ontology was also constructed in the mobile phone domain for their proposed system.

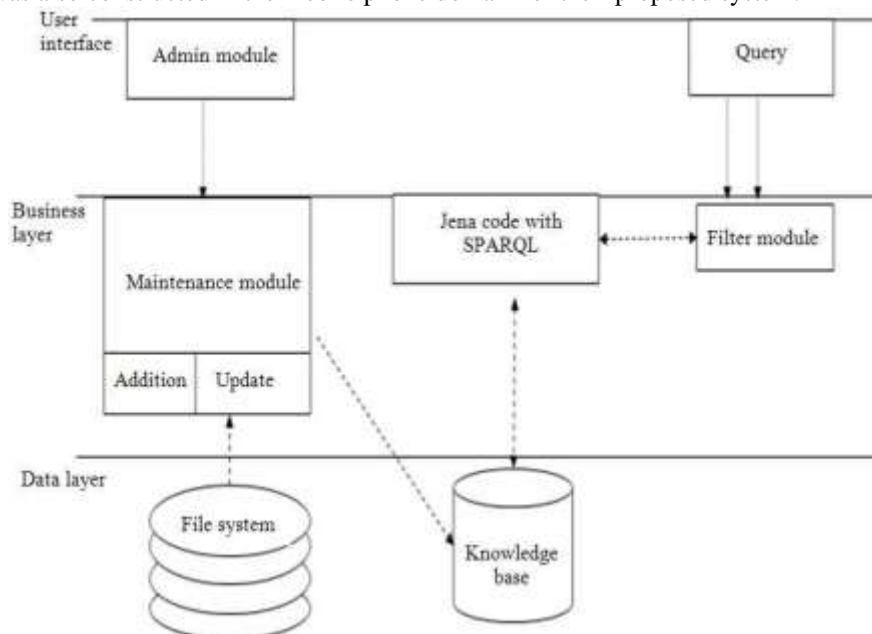


Fig. 1: Architecture of AKIMP

Additionally, a knowledge base for many applications can utilize the constructed ontology; such as a text classification, information acquiring and recommendation systems. In the near future, they would like to further automate the ontology construction by automatically learning the taxonomy part of ontology from text documents. To also consider the semantic facet of documents, different techniques of more competent concept extraction will be considered.

II. MATERIALS AND METHODS

A. Analysis

To form a basis for the system, Indian medicinal plants were chosen like Aloe Vera, Amla, Ashwagandha, Brahmi, Ginger, Turmeric, Neem, Peepal, Sandalwood, Tulsi. Information relevant to the diseases they can cure was searched from Wikipedia. The list of the diseases was obtained from TKDL.

B. Construction of Knowledge Base

Knowledge base is created for the domain of plants and their related diseases, extracting the data from Wikipedia and Traditional Knowledge Digital Library (TKDL). The information related to the herbal plants and their uses are extracted automatically by Java code. The extracted information is stored into Excel File in the form of Subject-Predicate-Object. This Excel file is automatically created by the Java code. The Excel in the form of triples (Subject-Predicate-Object) are verified with the domain experts before creating the Resource Description Framework (RDF)/Web Ontology Language (OWL).

The plant names along with the diseases were structured into classes. While constructing the classes, the existing plant ontology and disease ontology were considered as the reference. Plant names and disease names are identified as “Nouns” and those are compared with the existing hierarchy. Similarly the “Verbs” are identified along with the Noun Terms. From the unstructured text, the triples are constructed in the form of subject-predicate-object. These triples are stored in Excel format. Microsoft Excel tools are used to provide structure to all plants and diseases along with their relationships. These excel files are considered as the collection of triples for the reference so that they can be parsed to construct the ontology. The format of excel sheet used here is Microsoft Excel 97-2003 Worksheet (.xls).

C. Parsing

The data is parsed using java code into user interface using the jxl API. This API converts the data present in rows and column into text and stores in a String variable. We used this string to compare and show the output to the user. Sample code to read the triples from the excel file is shown in Fig 2.

```
String
plantname=Plants.plants.toLowerCase();
String address="E:/project/medicinal
plants/"+plantname+".xls";
public String readcolumn1 () throws
IOException
{
    String name="";
    try {
        Workbook work1 =
        Workbook.getWorkbook(new File(address));
        Sheet sheet1 = work1.getSheet(0);
        //String c1[]=new String[100];
        for (int i = 0; i < sheet1.getRows(); i++)
        {
            Cell a1 = sheet1.getCell(0, i);
            name+=a1.getContents().toString()+"\n";
        }
    } catch (BiffException e)
    {
```

Fig. 2: Sample Code to read the triples from Excel file

RDF, OWL(<http://www.w3.org/TR/owl-ref>) data forms are created using Jena, reading the input from Excel File. In Java, Ontology models are created through the Jena Model Factory. Jena is the Java enabled semantic web API that can be able to read and process the information from the knowledge base. Plants classification of plants is done based on botanical classification (Joyetal.,1998). Based on the existing classification, Jena searches for the corresponding class of the plant and the triple is created accordingly. Similarly for the available list of disease class the triple is created after identifying the proper class of the disease. The output of the Jena is in the form of RDF/OWL file that will be available for Protégé.

```
try {
disease=obj.readcolumn1();
link=obj.readcolumn2();
//image=obj.readcolumn3();
} catch (IOException e) {
e.printStackTrace();
}
textarea1.setText(verb+"\n"+disease);
textarea2.setText(link);
//url="www.google.com"/url by default
textarea2.addMouseListener(new MouseAdapter() {
@Override
public void mouseClicked(MouseEvent e) {
if (e.getButton() != MouseEvent.BUTTON1) {
return;
}
if (e.getClickCount() != 2) {
return;
}
int offset = textarea2.viewToModel(e.getPoint());
try {
int rowStart = Utilities.getRowStart(textarea2,
offset);
int rowEnd = Utilities.getRowEnd(textarea2, offset);
String selectedLine =
textarea2.getText().substring(rowStart, rowEnd);
url=selectedLine;
}
}
```

Fig. 3: Sample Code to display the subjects or objects

Protégé (Horridge et al., 2007) with the stored RDF/OWL is the knowledge base for further processing. Protégé is a free, open source ontology editor based on the Java platform. It is extensible, provides a plug-and-play environment, supports graphic visualization. Noy and McGuinness (2001) discussed about the ontology creation techniques using Protégé.

Defining classes in the ontology, arranging the classes in hierarchy: Classes are the main focus of most of the ontologies. A class can have subclasses that represent. Plantae Kingdom consists of Kingdom details, which is the subclass of "Thing" in Protégé. Order is the subclass of Kingdom, Family is the subclass of Order, Genus is the subclass of Order, Species is the subclass of Genus and Plant is the subclass of Genus. Similarly the disease ontology, classification is done. Since the details of Plants and Disease are mentioned in the form of text in the input sources (<http://www.tkd1.res.in/tkd1/langdefault/common/Home.asp?GL=Eng>; <http://en.wikipedia.org/wiki/Main-Page>).

OWL slots represent relationships among classes and instances. There are two main types of properties, Object properties and Data type properties. Object properties are relationships between two individuals. Object properties are used to relate two objects whereas Data type property used to relate one instance with any of the built-in data types. For example Object property "heals" is used to relate Plant instance and disease instance. Data property is another type of property that relates the instance with built-in data types and their values (Vadivu et al., 2011).

The application development of ontology based knowledge querying is made simple by using Jena programming toolkit. Class, property, individual creation is done using Protégé. Jena (<http://jena.sourceforge.net/>) aims to provide a consistent programming interface for ontology application development with the base of Java Programming. <<OntClass>> is used to represent OWL class or RDFS class. <<OntModel>> extends support to all the different types of objects expected to be in ontology: Individuals, properties (in the case of a property hierarchy) and classes (in the case of a class hierarchy).

In Java, Ontology models are created through the Jena Model Factory. O'Connor et al. (2007) discussed about the knowledge querying. SPARQL is a Simple Protocol and RDF Query Language. SPARQL uses pattern matching for querying RDF graphs. It is syntactically similar to SQL. SPARQL's features include basic consolidated patterns, value filters and optional

patterns. Thus using SPARQL in Jena it is possible to retrieve more specific and semantically related resources can identified without affecting the existing data models (Vadivu and Hopper, 2010).

III. RESULTS

Generating result window: The system enables the user to enter a disease/plant name and get the relevant plant/disease name and link to the Wikipedia page from the system. To take input from the user and generate result as per the requirement the Java windows builder (Swing Component) was used. The interface consists of an input field where input is given and this is then compared to the data stored in the previous step. If a match is found, the data is parsed and shown in the output field along with the Wikipedia link where user can get precise information. The programming that the system runs on is robust and does not need to be modified to expand the system. To supplement the system with the details of more plants, only the pertinent classes have to be added to it. This makes the approach very efficient.

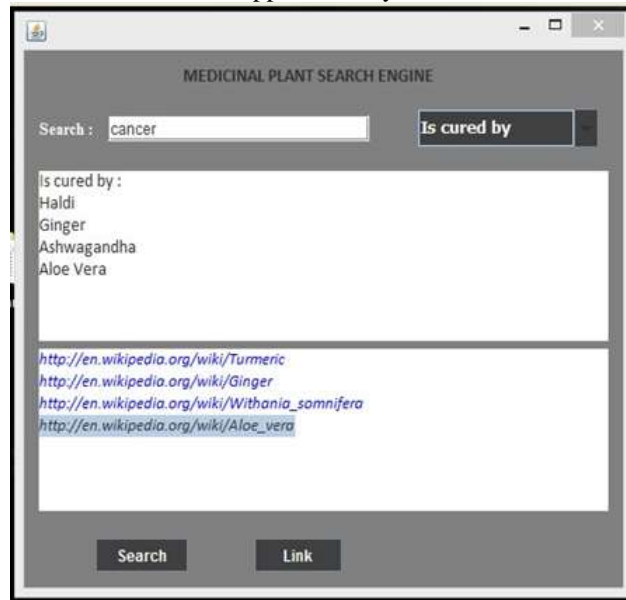


Fig. 4: Search by Disease

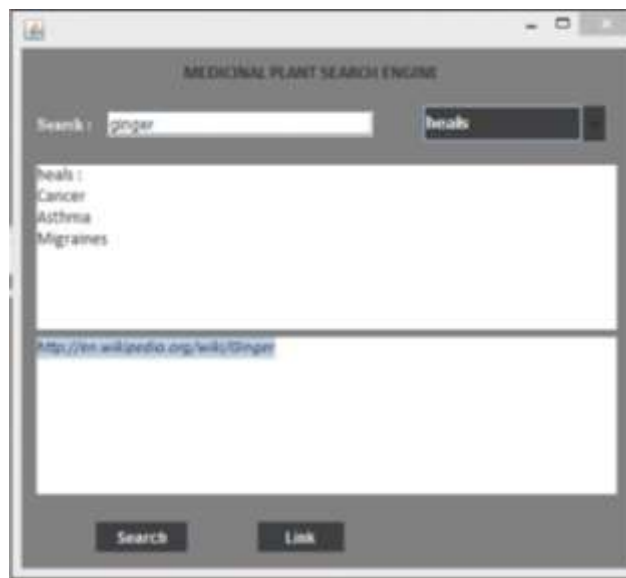


Fig. 4: Search by Medicinal Plant

IV. CONCLUSION

Thus it is possible to find the specific resultant web page based on the user requirement. This can be used for any domain specific web input pages with unstructured, structured, or semi structured data. This makes the existing search engines to find the more relevant web pages by introducing new layer of ontology. This ontology layer is constructed automatically.

REFERENCES

- [1] Mei-yingJia, et.al., Automatic OntologyConstruction Approaches and its Application on Military Intelligence, IEEE computer society, 2009.
- [2] Bao-An Nguyen, Don-Lin Yang., A Semi-Automatic Approach to Construct Vietnamese Ontology from Online Text, International review of research in open and distance learning, December 2012.
- [3] Vadivu.G, Waheeta Hopper, Ontology Mapping of Indian Medicinal Plants with Standardized Medical Terms, Journal of Computer Science, Science Publications, 2012. ISSN 1549-3636.
- [4] Antoniou,G. and F.V.Harmelen, 2004. A Semantic Web Primer.1stEdn., MIT Press, Cambridge, ISBN-10:0262012103,pp:258.
- [5] Azlida,M.,E.RahmanandA.Abd,2008.Organisingherbsknowledge: Isanontology ortaxonomythe answer?Proceedings oftheIEEEInternational SymposiumonInformationTechnology,Aug.26-28, IEEEExplorePress,KualaLumpur,Malaysia, pp:1-4.DOI:10.1109/ITSIM.2008.4631693
- [6] Barathi,M.,2011.Context disambiguation based semantic websearch for effective information retrieval. J. Comput. Sci., 7: 548-553. DOI: 10.3844/jcssp.2011.548.553
- [7] Berners-Lee,T., J.Hendlerand O.Lasilla, 2001. TheSemanticWeb.
- [8] Farooq, A.M., J.ArshadandA.Shah,2010.Alayered approach forsimilaritymeasurement between ontologies.J. Am.Sci.,6:69-77.
- [9] Fellbaum,C.,1998.WordNet:AnElectronicLexicalDatabase.1stEdn.,MITPress, Cambridge, ISBN-10:026206197X,pp:445.
- [10] Heim, P., S. LohmannandT. Stegemann, 2010. Interactive relationship discovery via the semantic web. Proceedings of the7th International Conference onTheSemanticWeb:ResearchandApplications, (ESWC'10), Springer-VerlagBerlin, Heidelberg,pp: 303-317.DOI:10.1007/978-3-642- 13486-9_21
- [11] Horridge,M.,S.Jupp,G.Moulton,A.RectorandR.Stevensetal., 2007. A PracticalGuideTo Building OWLOntologies UsingProt´eg´eandCO-ODE Tools.TheUniversityofManchester.
- [12] <http://www.tkd1.res.in/tkd1/langdefault/common/Home.asp?GL=Eng>