Journal of Computer Science 8 (9): 1576-1584, 2012
ISSN 1549-3636

Ontology Mapping of Indian Medicinal Plants with Standardized Medical Terms

[1]Vadivu, G. and [2]S. Waheeta Hopper
[1]Department of Information Technology,
[2]Department of Bio Informatics,
Sri Ramaswamy Memorial University, Kattankulathur, 603203, India

Abstract: Problem statement: World Wide Web (WWW) consisting large volume of information related with medicinal plants. However health care recommendation with Indian Medicinal Plants becomes complicated because valuable Information about medicinal resources as plants is scattered, in text form and unstructured. Search engines are not quite efficient and require excessive manual processing. Therefore search becomes difficult for the ordinary users to find the medicinal uses of herbal plants from the web. And another problem is that the domain experts could not able to map the medicinal uses of herbal plants with the existing standardized medical terms. Mapping the existing ontology introduces the problem of finding the similarity between the terms and relationships. Finding the solution to perform automatic mapping is another major challenge to be solved. **Approach:** To address these issues we developed a Knowledge framework for the Indian Medicinal Plants (KIMP). Knowledge framework includes the ontology creation, user 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). SPARQL Protocol and RDF Query Language (SPARQL) is used to retrieve various query patterns. Automated mapping is achieved by considering lexical and edge based relatedness. **Results:** The user interface is demonstrated for five thousand concepts, which gives the related information from Wikipedia web page in three languages. Mapping recommendation by the lexical similarity Jaccard algorithm gives 27% and Jaro Winkler algorithm gives 60%. Edge based relationship using WuPalmer algorithm gives 93% mapping recommendation. These are analyzed and compared with our algorithm based on WuPalmer gives more specific mapping results than WuPalmer with 71%. **Conclusion:** Thus it possible to find the specific resultant web page based on the user requirement in three different languages. The mapping with standardized ontology gives more improvement in analyzing the performance of the medicinal plants and their uses.

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

INTRODUCTION

India is the largest producer of medicinal herbs and is called the botanical garden of the world. India is blessed with rich and diverse heritage of cultural traditions. 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 difficult to search for the required information. Searching the specific information by the general user is a difficult process. Search engines are used to search for these documents, but they still have to be interpreted by themselves before any useful information could be extracted. And the text based herbal plant details are not mapped with the standardized medical terms which is required by the domain experts. As text based information, there are some limitations in using the medicinal plants:

- Searching text-based documents is very difficult
- They provide general information which is not more appropriate to the user need
- There is no mapping with the standardized medical terms

Corresponding Author: Vadivu, G., Department of Information and Technology, Sri Ramaswamy Memorial University, Kattankulathur, 603203, India

This study is used to address these limitations for providing useful information.

To cope with the existing web based problems with information searching the augmentation of meaningful contents in the web is a semantic based solution. Semantic Web was introduced by Berners-Lee *et al.* (2001). Semantic Web is an intelligent incarnation and advancement in World Wide Web to collect, manipulate and annotate the information by providing categorization, uniform access to resources and structuring the information in machine process able format. To structure the information in machine process able form, Semantic Web has introduced the concept of "Ontology" (Antoniou and Harmelen, 2004).

India possesses a rich traditional knowledge of ways and means practiced to treat diseases afflicting people. This knowledge has generally been passed down by word of mouth from generation to generation. A part of this knowledge has been described in ancient classical and other literature, often inaccessible to the common man and even when accessible rarely understood. Documentation of this existing knowledge, available in public domain, on various traditional systems of medicine has become imperative to safeguard the sovereignty of this traditional knowledge. References are also collected from Tamil (one of the regional language of India), English and Hindi (one of the regional language of India) Wikipedia related to medicinal plants (http://www.tkdl.res.in/tkdl/langdefault/common/Home.asp?GL=Eng; http://en.wikipedia.org/wiki/Main-Page).

Ontology describes the concepts, their relationships and properties within their domain and it can be utilized both to offer automatic inferring and interoperability between applications. This is an appropriate vision for knowledge management. Ontology provides understanding of the structure of information. With a common ontology, information that is spread out in many different applications and documents can be viewable in an easy way to understand and navigate. The ontology makes it possible to search both explicit and tacit knowledge, thereby bridging the gap between the tacit and explicit knowledge. The advantages of ontology are: knowledge sharing, logic inference and reuse of knowledge.

Ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

In practical terms, developing ontology includes:

- Defining classes in the ontology
- Arranging the classes in a taxonomic (subclass-superclass) hierarchy
- Defining properties (or slots)
- Filling in the values for properties of instances

Related works: Ontology based E-Health system with Thai Herb recommendation project is created the ontology for Thai herbs and based on the user input as symptoms, province of living, chronic disease details; the recommendations are given for treating the symptoms. But it is not considering the MeSH terms for treating the symptoms (Kato *et al*., 2010).

Designing a conceptual model for herbal research domain using ontology technique, discussed on how ontology technique can be used to represent conceptual model database design for herbal research domain (Mamat and Rahman, 2009).

The role of domain ontologies in database design: An ontology management and conceptual modeling environment, this study demonstrated how ontology representation can assist database design. Common ontology representation or basic relationships for conceptual modeling are-a, synonym and related-to. The purpose of this application is to simplify in defining the rules exist in herbal industry. The following four types of relationship component are Prerequisite, temporal, mutually inclusive and mutually exclusive are also explained (Sugumaran and Storey, 2006).

Organizing herbs knowledge: Is an ontology or taxonomy the answer? This study identified that ontology can be used to organize the information that have variety of concepts are need in sharing herb knowledge. Despite more problem solver pointed to ontology, taxonomy also important in identification and classification of herbs (Azlida *et al*., 2008).

A model driven ontology-based architecture for supporting the quality of services in pervasive telemedicine applications, discusses on ontology based architecture model enabling an intelligent pervasive telemedicine tasks management. Message exchange among different actors, the message exchanged by the system will be encapsulated in the XML format. For example, if the patient needs coronary angioplasty and need emergency physician to the closest hospital can be identified and exchanged as message (Nageba *et al.*, 2009).

The interactive aspect of relationship discovery, is dicussed in (Heim *et al*., 2010). The real discovery is only possible with a human involved, since only the user can ultimately decide if a found relationship is relevant in a certain situation or not.

A Methodology for Ontology Integration, ontology reuse is an important research issue only one of its sub processes is merging; the other reuse sub process is integration. In this study they described the activities that compose this process and describe a methodology to perform the ontology integration process (Pinto *et al*., 2004).

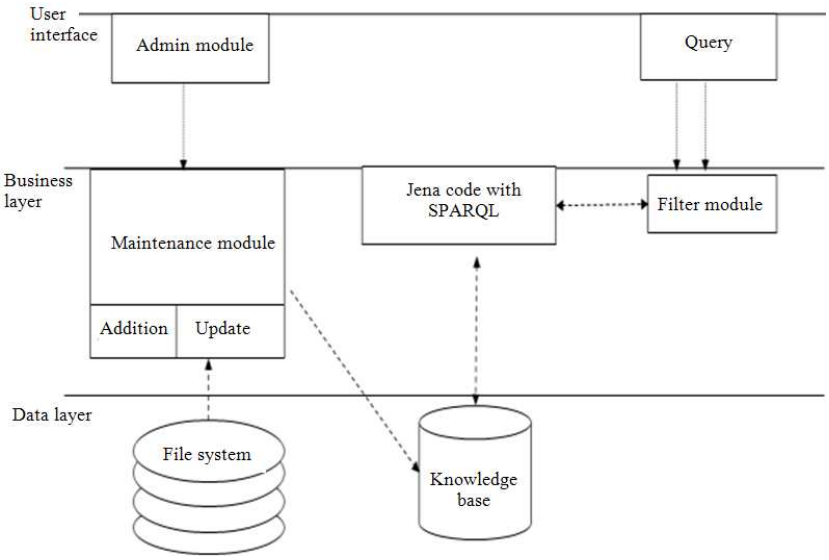

Fig. 1: Architecture of KIMP

MATERIALS AND METHODS

Construction of knowledge Base: Knowledge base is created for the domain of plants and their related disease, extracting the data from Wikipedia and Traditional Knowledge Digital Library (TKDL). RDF, OWL (http://www.w3.org/TR/owl-ref) data forms are created using Protégé (Horridge *et al*., 2007) and stored as the knowledge base for further processing. The protégé is a free, open source ontology editor based on java platform. It is extensible, provides a plug-and-play environment, support graphic visualization. Noy and McGuinness (2001) discussed about the ontology creation techniques using Protege. Jena is the Java enabled semantic web API framework which can able to read and process the information from the knowledge base.

Knowledge framework for Indian Medicinal Plants (KIMP) class classification of plants is done based on botanical classification (Joy *et al*., 1998). Disease terms mentioned in the KIMP ontology is mapped with the MeSH ontology automatically. User interface is created for the general users by giving the list of diseases and its corresponding properties available for those diseases. The overall architecture is shown in Fig. 1.

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 concepts that are more specific than the super class. Plant Kingdom consists of Kingdom details which is the subclass of thing. 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. Sample classification of the Plant classification is shown in Fig. 2. For the disease ontology, classification is not done at this time. Since the details of Plants and Disease are mentioned in the form of text in the input sources (http://www.tkdl.res.in/tkdl/langdefault/common/Home.asp?GL=Eng; http://en.wikipedia.org/wiki/Main-Page).

OWL Properties/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 instances whereas Data type property used to relate one instance with any of the built in data types. For example Object property usedToCure is used to relate Plant instance and disease instance. Data property is another type of property which 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 and its procedure is shown in Fig. 3. Class, property, individual creation is done using Protégé, which is shown in Fig. 4. 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 for the kinds of objects expected to be in ontology: Classes (in a class hierarchy), properties (in a property hierarchy) and individuals.

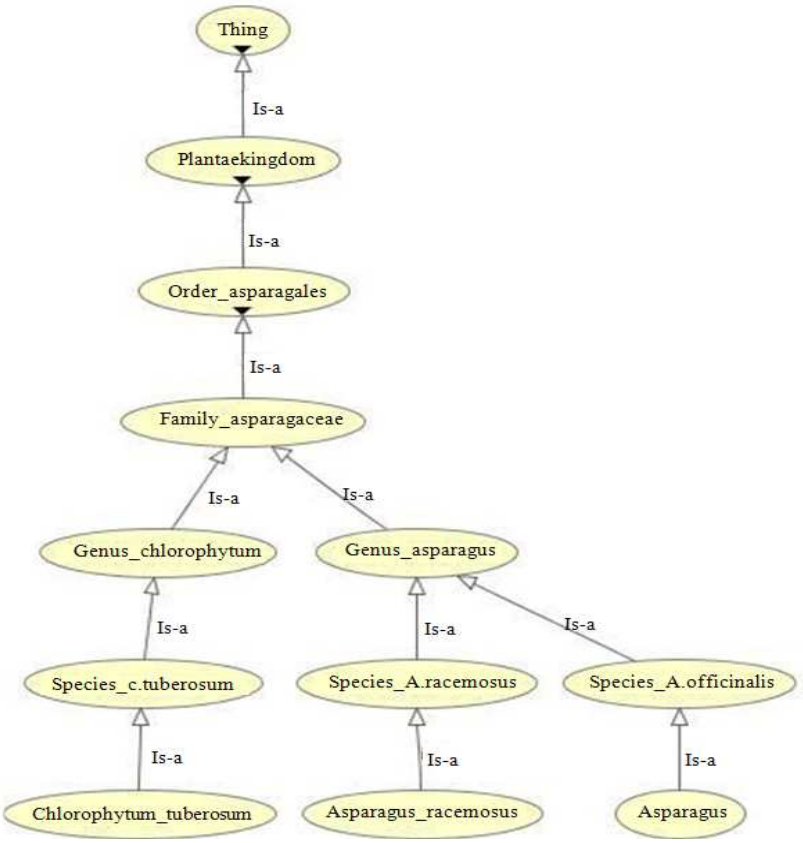

Fig. 2: Sample of Plant classification

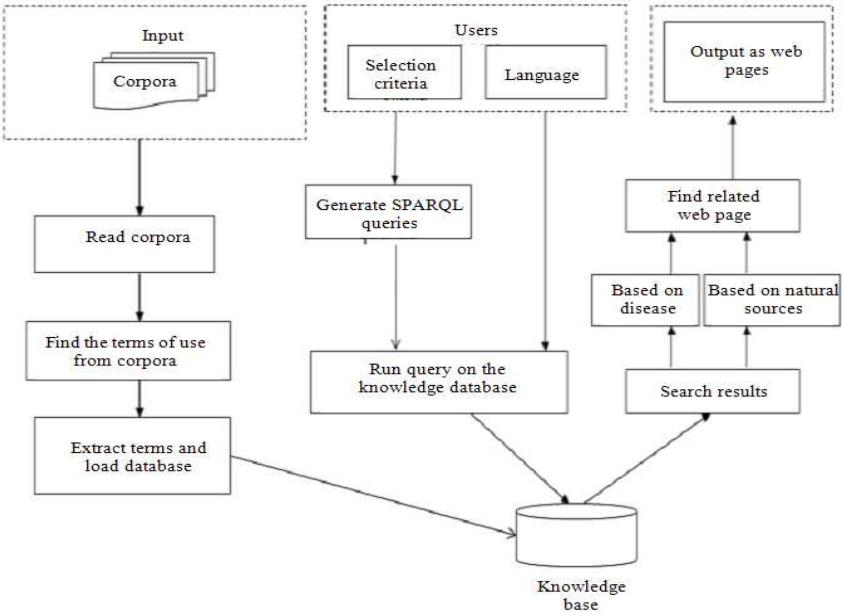

Fig. 3: Procedural diagram of KIMP

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 is a syntactically-SQL-like language for querying RDF graphs via pattern matching. The language's features include basic conjunctive 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).

MeSH, Medical Subject Heading, (http://www.nlm.nih.gov/pubs/factsheets/mesh.html) is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. Integrating this plant ontology and their medicinal uses with the existing Medical Subject Heading (MeSH) is useful to find more usage of the medicinal plants.

Mapping is one of the sub processes of integration which is the process of building ontology in one subject and reusing it by one or more other subjects. The steps of mapping process are to identify the available ontologies and then finding the possible terms to be mapped.

To find the terms to be mapped, semantic similarity between the ontology terms have to be calculated in automated way. Since for the large scale of data it is not possible to perform the manual mapping among the terms. Mapping of ontologies requires the class mapping, property mapping and instance mapping.

The following algorithm shows the mapping procedure.

Algorithm Map (O1, O2):
Input: KO (KIMP Domain Ontology), MO (MeSH ontology)

Output: Mapping recommendation between KO and MO.

1. Initialize set of values t ∈ C, t∈ P, t ∈ I. C-class, P- Property, I–Instance/Individual.
2. Repeat
3. Select values from C, P, I
4. Let G, G' from KO and MO
5. For (t, t')∈ G ×G' do
    a. Compute similarity of t, t'.
    b. Choose the highest similarity value of t, t'
    c. Add the mapping of m(t, t') into M
6. end for
7. Until no more values available.
8. Return M.

Similarity measures: The similarity measuring methods are discussed in (Farooq *et al*., 2010) Similarity measure between classes, properties and individuals is used to find the mapping between the terms. In this study, we have implemented lexical and edge based counting measures.

Lexical algorithms are based on the string matching algorithm. We have used Jaccard (http://en.wikipedia.org/wiki/Jaccard-index) and JaroWinkler, another lexical based (http://alias-i.com/lingpipe/docs/api/com/aliasi/spell/JaroWinklerDistance.html) algorithm to find the lexical similarity between KIMP ontology and MeSH ontology terms. Wordnet, (Fellbaum,1998) database is used as the base database for finding the similarity score. Word Net is a large lexical database of English.

Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser.

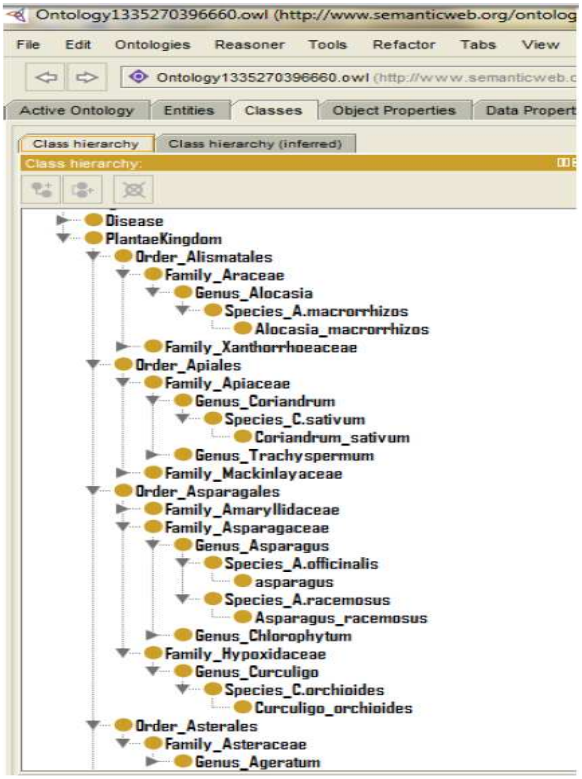

Fig. 4: Part of KIMP Class hierarchy

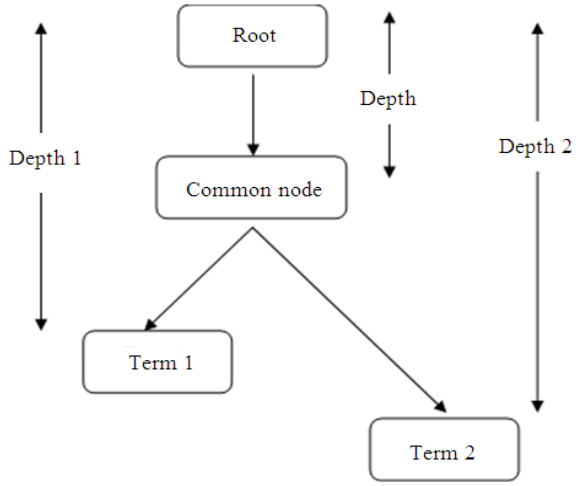

Fig. 5: Values of depth1, depth2 based on Wu and Palmer algorithm

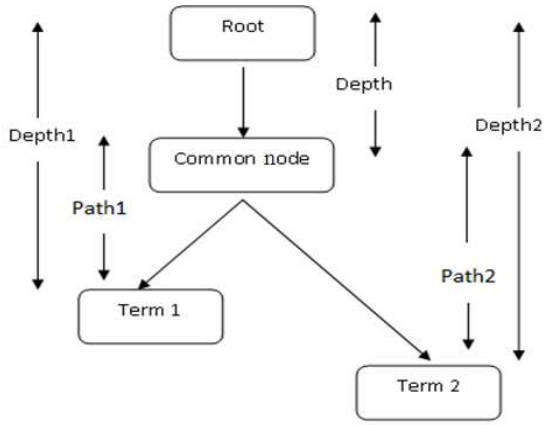

Fig. 6: Modified diagram of WuPalmer

The Jaccard coefficient measures similarity between the words A and B and is defined as the size of the intersection divided by the size of the union of the sample words A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Distance values calculated between 0 and 1, distance of 0 means, the character sequences share all of their terms, whereas a distance of 1 means they have no characters in common. The following is the code for Jaccard distance which will return the values between 0 and 1 based on the string similarity:

```
For (String x: s1)
  if (s2.contains(x))
    ++numMatch;
  int numTotal = s1. Size () + s2. Size ()-numMatch;
  return ((double) numMatch)/((double) numTotal);
```

Jaro and Winkler lexical similarity algorithm is also used for the same purpose. Based on Jaro, the distance d_j of two given strings s_1 and s_2 is:

$$d_j = 1/3[(m/|s_1|) + (m/|s_2|) +((m-t)/|s_1|)]$$

where: m is the number of matching characters; t is half the number of transpositions. (Wu and Palmer, 1994) distance uses a prefix scale p which gives more favourable ratings to strings that match from the beginning for a set prefix length l. Given two strings s1 and s2, their Jaro-Winkler distance d_w is:

$$d_w = d_j + (l_p(1- d_j))$$

Where:

d_j = The Jaro distance for strings s_1 and s_2

l = The length of common prefix at the start of the string up to a maximum of 4 characters

p = A constant scaling factor for how much the score is adjusted upwards for having common prefixes. P should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is p = 0.1

```
double weight = (numCommonD/len1
                + numCommonD/len2
                +(numCommo-
numTransposed)/numCommonD)/3.0;
```

Distance values calculated between 0 and 1, distance of 0 means the character sequences share all of their terms, whereas a distance of 1 means they have no terms in common.

Both Jaccard and JaroWinkler algorithms are used to find the lexical similarity between the string and conceptual similarity measure is not included for improving the mapping.

Edge based counting algorithm is used to find conceptual relationship among the terms. We have used Wu and Palmer (Wu and Palmer, 1994) algorithm as the basic to find edge based algorithm and the related diagram is shown in Fig. 5.

```
int depth1 = depthFinder.getShortestDepth
             (synset1);
int depth2 = depthFinder.getShortestDepth
             (synset2);
```

```
double score = 0;
if (depth1>0 and depth2 >0) {
  score = (double)( 2 * depth )/
        (double)( depth1 + depth2);}
```

The above code is based on WuPalmer algorithm. We analyzed WuPalmer algorithm and identified that Wu and Palmer algorithm does not give more accurate values because it always considers the depth of the terms from the root node. Calculating the edge distance from the common node from where the terms are getting divided into different paths will give better results. Based on this we have developed KIMP_WuPalmer algorithm which gives more accurate similarity values than WuPalmer.

The following code is the modified version based on WuPalmer algorithm. Fig. 6 shows the modified concept of Wu and Palmer.

```
int depth1 = depth Finder. Get Shortest Depth
             (synset1);
int depth2 = depth Finder. Get Shortest Depth
             (synset2);
double score = 0,path1=0,path2=0, path_dist=0;
if (depth1>0 && depth2 >0) {
    path1=depth1-depth;
    path2 = depth2-depth;
    path_dist=path1+path2;
  score = ((double)( 2 * depth ) /
 (double)(depth1 +depth2))*(1.0/path_dist);
    }
```

RESULTS

Barathi (2011) also discussed about the diaambiguation of user queries. Naïve users can retrieve their required information by selecting the plant name or disease name. After selecting this, the associated properties will be listed in the list box and in which language the user wants to view the result. The output will be the specific required web page from Wikipedia or from TKDL, shown in Fig. 7.

The sample mapping recommendation of Jaccard lexical based measure is shown below:

Jaccard Mapping of KIMP ontology with with MeSH ontology:

abdominal_lump abdominal_absces
Jaccard Lexical Distance 0.5
abdominal_lump abdominal_aortic_aneurysm
Jaccard Lexical Distance 0.6
abdominal_lump abdominal_fibromatosis
Jaccard Lexical Distance 0.5
abdominal_lump abdominal_hernia
Jaccard Lexical Distance 0.5
abdominal_lump abdominal_neoplasm Jaccard Lexical Distance 0.5
abdominal_lump abdominal_pregnancy Jaccard Lexical Distance 0.5
heart_disease abducens_nerve_disease Jaccard Lexical Distance 0.6
intermittent_fever abietane_diterpene Jaccard Lexical Distance 0.75

The sample mapping recommendation of Jaro Winkler lexical based measure is shown below:

JaroWinkler mapping of KIMP ontology with with MeSH ontology:

abdominal_lump abdominal_absces
Jaro Winkler Lexical Distance 0.13214285714285712
abdominal_lump abdominal_aortic_aneurysm
Jaro Winkler Lexical Distance 0.1548571428571428
abdominal_lump abdominal_fibromatosis
Jaro Winkler Lexical Distance
0.1428571428571429
abdominal_lump abdominal_hernia
Jaro Winkler Lexical Distance
0.13214285714285712
abdominal_lump abdominal_neoplasm
Jaro Winkler Lexical Distance
0.08522588522588526
abdominal_lump abdominal_pregnancy
Jaro Winkler Lexical Distance
0.12706766917293233

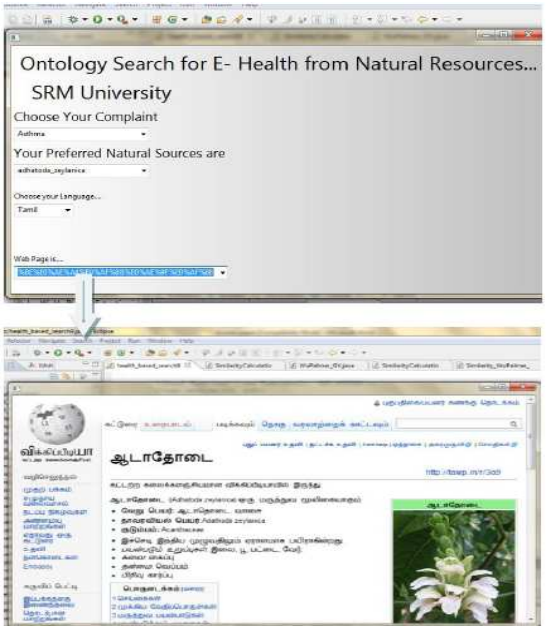

Fig. 7: Searching from KIMP

DISCUSSION

The result of Jaccard mapping values were analyzed with different threshold values and verified manually. This gives 27.81% of mapping recommendation. The results are obtained by Jaro_Winkler were analyzed with different threshold values for similarity measure and verified manually. This gives 60.96% of mapping recommendation.

Based on Wu and Palmer the more similar words are identified based on the hierarchical structure of the MeSH ontology. 93% of the terms are mapped based on Wu Palmer algorithm. Comparison of Jaccard, Jaro Winkler and Wu Palmer is shown in Fig. 8.

Figure 9 shows the comparative results of Wu Palmer and KIMP_WuPalmer and KIMP_Wu Palmer result gives more accurate results than Wu and Palmer (1994).

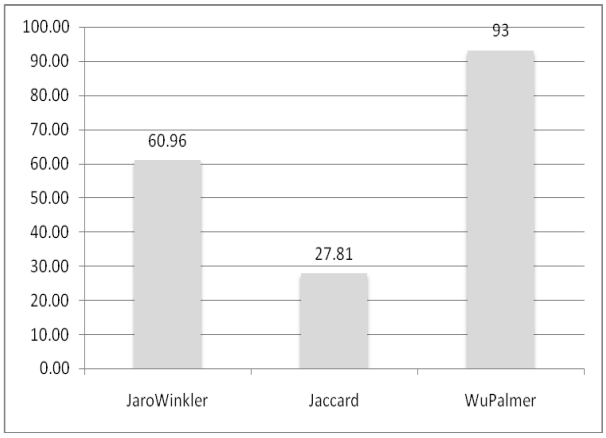

Fig. 8: Result analysis of Jaccard, Jaro-Winkler and WuPalmer Algorithms

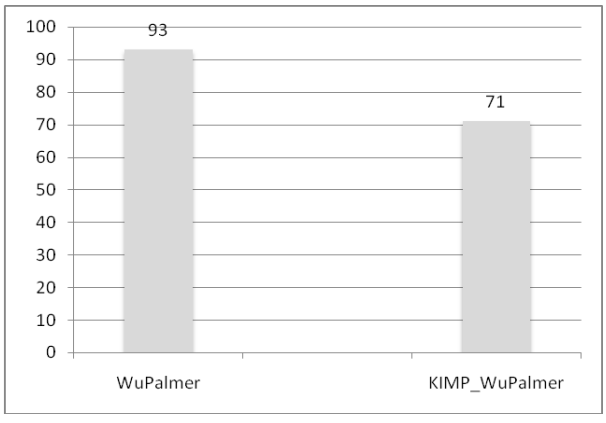

Fig. 9: Comparison of WuPalmer and modified KIMP_WuPalmer

CONCLUSION

Thus it possible to find the specific resultant web page based on the user requirement in three different languages. Jaccard, Jaro Winkler algoritms are used to find the lexical similarity which considers only the string matching. Wu and Palmer (1994) consider the edges between the terms to find more conceptual relationship which gives more related terms. Our algorithm based on WuPalmer considers the depth of the terms with more appropriate value to find better results. The mapping with standardized ontology will be useful in analyzing and improving in identifying the uses of medicinal plants.

REFERENCES

Antoniou, G. and F.V. Harmelen, 2004. A Semantic Web Primer. 1st Edn., MIT Press, Cambridge, ISBN-10: 0262012103, pp: 258.

Azlida, M., E. Rahman and A. Abd, 2008. Organising herbs knowledge: Is an ontology or taxonomy the answer? Proceedings of the IEEE International Symposium on Information Technology, Aug. 26-28, IEEE Xplore Press, Kuala Lumpur, Malaysia, pp: 1-4. DOI: 10.1109/ITSIM.2008.4631693

Barathi, M., 2011. Context disambiguation based semantic web search for effective information retrieval. J. Comput. Sci., 7: 548-553. DOI: 10.3844/jcssp.2011.548.553

Berners-Lee, T., J. Hendler and O. Lasilla, 2001. The Semantic Web.

Farooq, A. M., J. Arshad and A. Shah, 2010. A layered approach for similarity measurement between ontologies. J. Am. Sci., 6: 69-77.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. 1st Edn., MIT Press, Cambridge, ISBN-10: 026206197X, pp: 445.

Heim, P., S. Lohmann and T. Stegemann, 2010. Interactive relationship discovery via the semantic web. Proceedings of the 7th International Conference on The Semantic Web: Research and Applications, (ESWC' 10), Springer-Verlag Berlin, Heidelberg, pp: 303-317. DOI: 10.1007/978-3-642-13486-9_21

Horridge, M., S. Jupp, G. Moulton, A. Rector and R. Stevens *et al*., 2007. A Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 and CO-ODE Tools. The University of Manchester.

Joy, P.P., J. Thomas, S. Mathew, B.P. Skaria, M. Plants, 1998. Aromatic and medicinal plants research station. Kerala Agricultural University.

Kato, T., N. Maneerat, R. Varakulsiripunth, S. Izumi and H. Takahashi *et al*., 2010. Provision of Thai herbal recommendation based on an ontology. Proceedings of the 3rd Conference on Human System Interactions (HSI), May 13-15, IEEE Xplore Press, Rzeszow, pp: 217-222. DOI: 10.1109/HSI.2010.5514565

Mamat, A. and A.A. Rahman, 2009. Designing a conceptual model for herbal research domain using ontology technique. Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, Nov. 30-Dec. 2, IEEE Xplore Press, Pisa, pp: 1167-1172. DOI: 10.1109/ISDA.2009.192

Nageba, E., J. Fayn and P. Rubel, 2009. A model driven ontology-based architecture for supporting the quality of services in pervasive telemedicine applications. Proceedings of the IEEE 3rd International Conference on Pervasive Computing Technologies for Healthcare, Apr. 1-3, IEEE Xplore Press, London, pp: 1-8. DOI: 10.4108/ICST.PERVASIVEHEALTH2009.5968

Noy, N.F. and D.L. McGuinness, 2001. Ontology development 101: A guide to creating your first ontology. Development, 32: 1-25.

O'Connor, M., R. Shankar, S. Tu, C. Nyulas and D. Parrish *et al*., 2007. Using semantic web technologies for knowledge-driven querying of biomedical data. Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIM' 07), Springer-Verlag Berlin, Heidelberg, pp: 267-276. DOI: 10.1007/978-3-540-73599-1_36

Pinto, H.S. and J.P. Martins, 2004. Ontologies: How can they be built? Knowl. Inform. Syst., 6: 441-464. DOI: 10.1007/s10115-003-0138-1

Sugumaran, V. and V.C. Storey, 2006. The role of domain ontologies in database design: An ontology management and conceptual modeling environment. ACM Trans. Database Syst., 31: 1064-1094. DOI: 10.1145/1166074.1166083

Vadivu, G. and S.W. Hopper, 2010. Semantic linking and querying of natural food. Int. J. Comput. Appli., 11: 55-38. DOI: 10.5120/1567-2093

Vadivu, G., S.W. Hopper and G. BharathRam, 2011. Semantic data integration and querying using SWRL in LNCS. Springer-Verlag.

Wu, Z. and M. Palmer, 1994. Verbs semantics and lexical selection. Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, (ACL' 94), ACM Press, USA., pp: 133-138. DOI: 10.3115/981732.981751