# A Novel Approach For Semantic Similarity Measurement Using Spice Ontology And Matchmaking

Neema Babu, Fr. Rubin Thottupuram

*PG scholar, Amal Jyothi College of Engineering, Kanjirappally
**Faculty, Amal Jyothi College of Engineering, Kanjirappally

## Abstract

*Information Extraction aims to retrieve certain types of information from natural language text by processing them automatically. For example, an information extraction system might retrieve information about commercial spices in a country from a set of web pages while ignoring other types of information. Ontology-based information extraction has recently emerged as a subfield of information extraction. In this paper, we provide an ontology-based information extraction for spices especially for Black Pepper. The general idea in this paper is to employ a semantic annotation technique and similarity measurement approach by using the spice ontology for semantic information extraction. The present work uses a spice ontology that can be updated by training data set and the annotation process. We propose a framework that takes semi–structure documents from different resources and semantically annotates them. Then, a matchmaker system investigates similarity between a user's needs and meta data provided by the annotation.*

*Keywords: Semantic Web, Information Extraction, Black Pepper Ontology, Matchmaking*

## 1. Introduction

As the information on the Internet dramatically increases, more and more limitations in information searching are revealed, because web pages are designed for human use by mixing content with presentation. The problem with the present web includes the difficulty of retrieving documents, extracting relevant data from retrieved documents and combining information from different sources to achieve a particular goal. In order to overcome these limitations, the Semantic Web, based on ontology, was introduced by W3C to bring about significant advancement in web searching. To enable efficient web searching, ontology based semantic similarity measurements can be used. This paper proposes spice ontology which is mainly concentrating on Black pepper diseases.

India being the land of spices has the competitive advantage over a period of time to spices for its intensified quality. The black pepper is an important inter-crop in coconut and arecanut gardens which provides additional income to the farmers. However, with a significant higher production level, this needs atmost care in risk management starting from the choice of cultivar, depending upon agro-climatic zones, management practices including the management of nutrients, light, water, harvesting and disease control. If the black pepper is not handled scientifically after the harvest, it is developed afro toxins, which becomes impediment in marketing of the produce. Evidently, to harness the best potential of black pepper, the information about the black pepper diseases and it's control, which can become the guiding tool to the farmer, is essentially desirable. The main category which is related with Black Pepper ontology we consider here is: Black pepper Diseases.

To effectively find the best similar web documents, it is important to use semantic technology. The use of semantic descriptions of black pepper diseases allows for matching to improve the process of finding required black pepper diseases information farmers need. The next sections are dealing about the basics of semantic web technology and a brief description of our proposed work.

## 1. Semantic Web

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The Semantic Web technologies and standards are presented in figure below, where the layered representation of the technologies can be found.

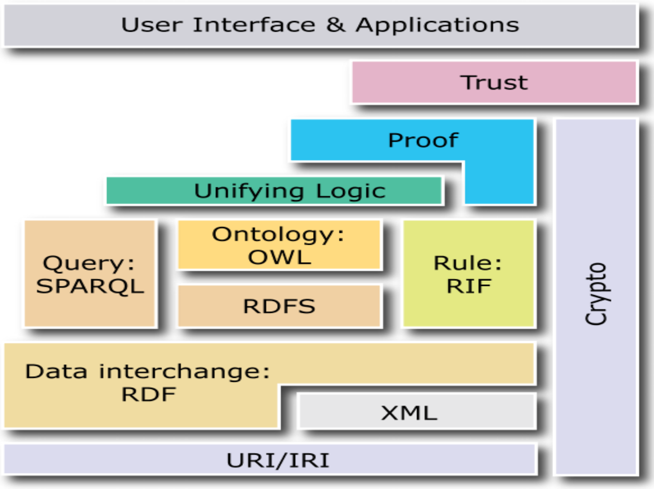

Fig1: Semantic Layers

If one takes a closer look at the semantic web layer cake, there is the "URI/IRI" Layer at the bottom of the figure, which represents the basic technology on which the Semantic Web is build on, the Uniform Resource Identifier and Internationalized Resource Identifier. This layer enables the resources defined at higher layers to be identified unique with help of World Wide Web identifiers. Also this layer represents the connection between the resources and the World Wide Web, where the later acts as a storage engine and information holder.

The next higher layer represents the "language" of the Semantic Web, used because of its popularity and simplicity - XML. XML enables the description of Semantic Web resources in a machine and human readable format. The "XML Query" layer represents the ability to search through XML resources, using different XML-based technologies like XPath and XML Schema. The other layers represent the Semantic Web Technologies, which have been recommended by W3C. RDF and RDFS represent the basic Semantic Web Technologies. They base on the XML format and can be used to build basic semantic web resources with basic relations. The Ontology layer represents the OWL - Web Ontology Language. It extends the RDFS Layer with three different extension languages (OWL Lite, OWL DL and OWL Full) and acts as a framework to enable the creation of complex ontologies. To get a high level query ability for ontologies, the "Rules / Query" Layer is used.

## 2.1 Ontology

The term Ontology has its roots in the philosophical domain. In order to understand the basic structure of our world and the study of existence, the word ontology has been connected with a branch of metaphysics. The problem is that the philosophical definition of ontology is not easy to port to the scientific domain. Therefore Dunwoodie uses an intelligible definition of ontology: "An ontology is a detailed model/picture/schema of a slice of reality which is based on the facts that we know about that reality. This model/picture/schema is a description of some of the things and some of the relationships between the things that are known about that reality". Helfin, defines the term "Ontology" as follows:"An ontology defines the terms used to describe and represent an area of knowledge". These ontologies can be shared by different applications, people and databases within a domain. A domain can be an area of knowledge, like medicine or a specific subject area. The definitions of ontologies are machine readable and they describe basic concepts in the domain and the relations between them. The knowledge, which is encoded in ontologies, is reusable due to the fact that the encoded knowledge can span different domains. Ontologies are able to specify the following kinds of concepts, which enable the description of almost every knowledge:

- Classes (things)
- Relationships between things
- Properties (attributes) of things

There are many motivations for developing and using ontologies:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

To enable efficient web searching, ontology based semantic similarity measurements can be used. Our work proposes a spice ontology which is mainly concentrating on Black pepper diseases.

## 2. Proposed Work

The Internet provides us with enormous amounts of data. The problem about this large amount of data is how to extract the information from this data and how to enable machines to collect process and interpret this data automatically. The main task of the Semantic Web should be that the Browser should browse for us, in a way of understanding our needs and taking advantage of the semantic technologies.

There is a large amount of data in the World Wide Web, but the current XHTML standard does not provide the possibility to add semantic meaning to the data. The information is readable and interpretable only by human readers and cannot be interpreted automatically by machines. As an example, looking at the Wikipedia page provides a reader with enormous amounts of information, but only for human readers, not for a machine. This means that the information is only data for the machine, which cannot be interpreted or used for further processing.

Another example is the popular search engine Google, where one can query any search phrase. In most cases, the search will result in more than one million search results. In particular the search phrase ”Barack Obama” returns a search result set of 164000000 entries in Google and 309000 results in bing, which is impossible to browse for any user.

These are some problems associated with normal web searching. In our work we are considering an ontology based information extraction system which will reduce the above mentioned problems upto some level. We are concentrating in the area of diseases that affects the Black Pepper, which is a largely grown spice in our country. Our work will improve the search results related to that domain.

The present work uses a spice ontology that can be updated by training data set and the annotation process. We propose a framework that takes semi–structure documents from different resources and semantically annotates them. Then, a matchmaker system investigates similarity between a user’s needs and meta data provided by the annotation.

## 3.1 Overall Architecture

The following figure shows the overall architecture of the proposed system. It involves collection of semi-structure documents from different sources, conversion of that collected documents to OWL (Web Ontology Language) format and the creation of spice ontology and finally the comparison of OWL documents with spice ontology. This will produce the ultimate search result.

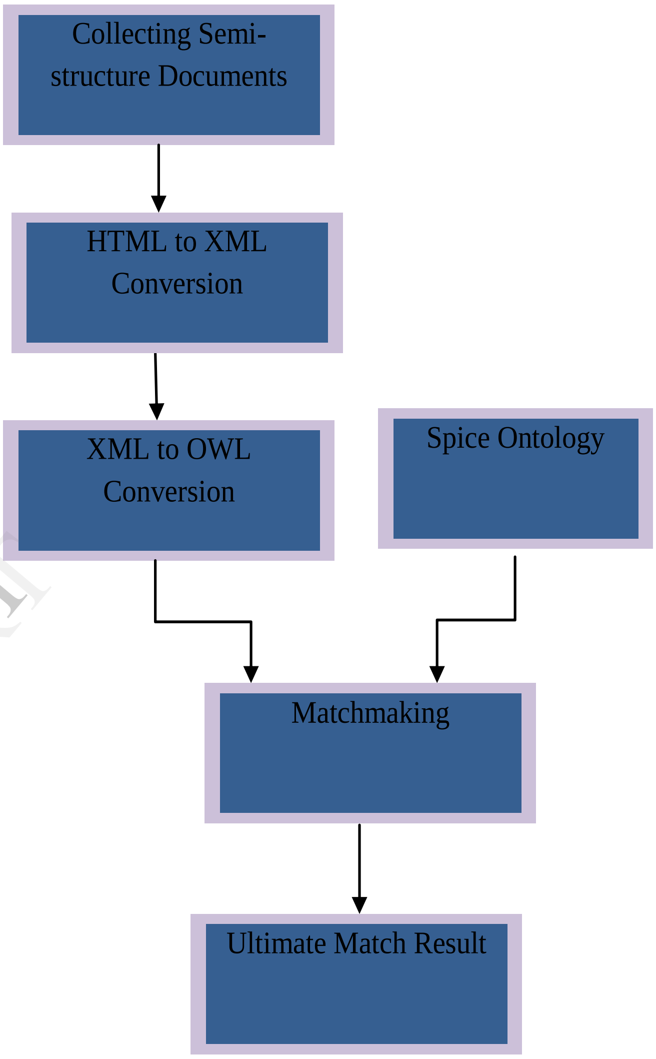

Fig2: Proposed System Architecture

### 3.1.1 Spice Ontology Creation

In our work the first step is the creation of spice ontology. Before that we have to consider the scope of this work. So the following sections are dealing with the scope of this work and the technologies used for the creation of spice ontology.

India being the land of spices has the competitive advantage over a period of time to spices for its intensified quality. Among the species, black pepper (Piper nigrum L), which is largely grown in Kerala, is an important spice in domestic and overseas market. If the black pepper is not handled scientifically after the harvest, it is developed afro toxins, which becomes impediment in marketing of the produce. The main category which is related with Black Pepper ontology we consider here is: Diseases.

The scope of this project is to infer new knowledge in the area of black pepper diseases and it supports scientists in farm-agro area. The practice improves the reliability and stability of information searching of black pepper diseases.

The software tool we have used for the creation of spice ontology is Protégé 4.3. Protégé is a free, open-source platform to construct domain models and knowledge-based applications with ontologies. Ontologies are now central to many applications such as scientific knowledge portals, information management and integration systems, electronic commerce and web services.

OWL ontology consists of different components like Classes, Properties and Individuals. With the help of these components it is possible to represent the concepts in a simpler way.

OWL classes are interpreted as sets that contain individuals. Classes may be organized into a superclass-subclass hierarchy, which is also known as taxonomy. Subclasses specialize (`are subsumed by') their superclasses. For example in our work consider the classes Spices and Pepper (Pepper might be a subclass of Spices (so Spices is the super class of Pepper). This says that, `All peppers are spices'. Some of the classes in our work are Black pepper, Diseases, Symptoms, Pests and Control.

Properties are binary relations on individuals - i.e. properties link two individuals together. Properties are of two types and that are

- Object Properties and
- Data Properties

For example in our ontology 'canbeAffectedBy' is an object property that relates the classes Black pepper and Diseases. Similarly 'hasSymptom' is the data property of the class Symptom.

Individuals, represent objects in the domain in which we are interested. For our work Black pepper varieties are the individuals of the class Black pepper. For example 'Arakkulammunda' is a variety of Black pepper and it is the individual of the class Black pepper. Similarly for other classes also we can define individuals.

By defining these components ontologies are used to capture knowledge about some domain of interest. Furthermore, the logical model allows the use of a reasoner which can check whether or not all of the statements and definitions in the ontology are mutually consistent and can also recognize which concepts fit under which definitions. The reasoner can therefore help to maintain the hierarchy correctly.

The following is the snapshot of our Black pepper disease ontology.

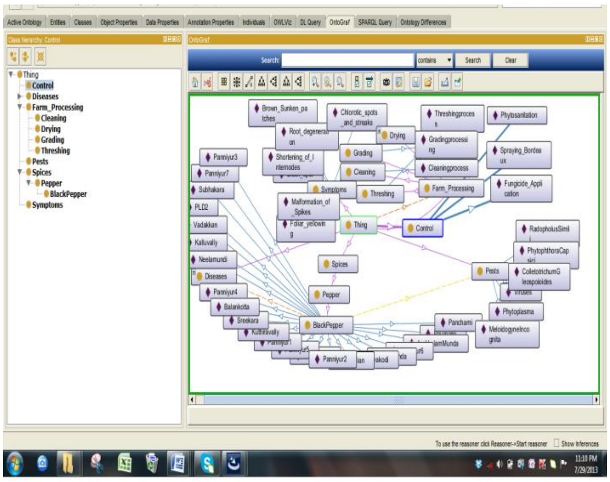

Fig 3: Black pepper Disease Ontology Graphical Representation

## 3.1.2 Matchmaking

Now our ontology is created. Next step is the matchmaking of semi-structure documents with the ontology and it involves the following steps:

- HTML to XML conversion
- XML to OWL (Web Ontology Language) conversion

For the conversion of html documents to xml, tools are available. Either we can use java code for this conversion. Similarly we have to convert the xml formatted document to OWL/RDF format. Otherwise we can’t compare it with the ontology.

In the present work, we take advantage of WordNet as a knowledge resource in order to find similarity between information objects (classes and instances) in the text corpus. WordNet is beyond of a just lexicon for English language. It contains hierarchies of words that organize words in certain hierarchies with certain relations. For each word, a set of synonym words exist that is called synset. There are various relations between synsets in the WordNet hierarchies such as hyponymy and coordinate.

The similarity of two words is a correlated concept with the semantic distance. It means that the more similarity exists between two words the less semantic distance they have. The easiest description of the semantic similarity is based on the number of edges that a path meets between two concepts in a WordNet hierarchy. More accurate definition of Semantic Similarity (SS) is:

$$sim(c_1, c_2) = -log \frac{len(c_1,c_2)}{2D},$$

−where,:

- $c_1$, $c_2$ are two synsets in the WorNet, and len($c_1$, $c_2$ ) refers to the number of nodes would be met in a path from $c_1$ to $c_2$;
- D denotes to the overall depth of the taxonomy.

A match happens between two synsets $c_1$, $c_2$ from recognized tokens of extracted information and instances belong to classes of designed ontology when:

$$sim(c_1, c_2) \geq \beta,$$

where: β is a constant called Similarity Constant. In our work, iff β equals to 1, it means that we would like to find exact same or very similar matches between keywords in the documents and ontology instances. In other cases, a routine greedy algorithm can determine β based on expected time and matching coverage of the matchmaker implementation.

## 4. Result and Conclusion

In the present paper, we proposed a framework to build a matchmaking system that uses the concept of semantic annotation and semantic similarity in order to find similar document with the user's needs among existing resources. We took advantages from the concept of WordNet based similarity and ontology(based similarity to evaluate amount of similarity between documents. We combined these two semantic similarity approaches with the suggestion of ontology updating algorithm. By proposing some similarity metrics that considers the importance of each section of the document, we are able to compare a part or whole of documents. Although our approach returns satisfying performance, it suffers from some short comes, in particular, interactions with the user of the system. Scalable systems that can process enormous amount of documents, should work automatically with the least intervention by the human user. Another short come is our limited training and test data set. Document corpus should be extended to have enough test training and test data. Future work has been planned to solve these problems.

## 5. References

1. .Alireza Ensan, IEEE, Student Member, Yevgen Biletskiy, IEEE, Member "Matchmaking Through Semantic Annotation and Similarity Measurement"
2. Y. Biletskiy, J. Anthony Brown and G.R. Ranganathan. "Information extraction from syllabi for academic e-Advising."
3. L C. Leacock and M. Chodorow, Combining local context and WordNet similarity for word sense identification.
4. Qian Gao "Similarity Matching Algorithm For Ontology-based Semantic Information Retrieval"
5. P. Resnik., "Using information content to evaluate semantic similarity"
6. www.semanticweb.org
7. www.w3.org
8. WordNet: http://wordnet.princeton.edu/,
9. http://www.spices.res.in/
10. www.kau.edu (Kerala Agricultural University Official Site)
11. The Package of Practices (Crops) (2011 Edition)