Novel Design of a Domain Specific Search Engine using Ontology

Venu Rathi
M.Tech Scholar
Shobhit University, Meerut, India

Niraj Singhal, Ph.D
Associate Professor
Shobhit University, Meerut, India

ABSTRACT
A web search engine is called domain specific if it is intuitive that some words or terms are more important than others. These are words specific to a domain. They should carry different weight which will affect the result of domain specific search engine. The objective of the proposed work is to highlight the property of ontology and to overcome the drawbacks of domain specific search engines using ontology. This work presents the design, development and implementation of a domain specific search engine. It also presents methods to compute relevance and weights of words using Ontology.

Keywords
Domain Specific Search Engine, Ontology, relevance score, syntable, weight table.

1. INTRODUCTION
A web search engine is a document retrieval system which helps find information stored in a computer system, corporate or proprietary network, or in a personal computer. It mainly searches for the documents in the World Wide Web (web), present in Hypertext Markup Language (HTML) format. HTML is unable to provide description as well as meaning of data. Current search mechanisms make use of syntax-based mechanisms in which the search and retrieval is based on the keywords entered by the user. Although users still often face with the daunting task of sifting through multiple pages of result, many of which are irrelevant. This is very time consuming task.

Search engines are web services that can be used to search and retrieve information. Web services are application components based on open standards like Extensible Markup Language (XML), Web Services Description Language (WSDL), Universal Description, Discovery and Integration (UDDI), and can be used by other services or users to communicate over the Internet. In web service, the search mechanism is keyword-based, hence inefficient [6]. The efficiency of matchmaking can be improved by the concept of ontology. The term ontology [1] is an old term used in the field of knowledge representation, information modeling, etc.

This paper presents design and implementation of a domain specific search engine using ontology. It identifies web page contents and calculates relevance score for that page, in order to select relevant web page to specific search engine as the results returned are relevant pages in response to a user query.

2. RELATED WORK
A web search engine is able to read every searchable page on the web, create an index of the information it finds and stores in a database. It compares that information with user's search request and return results back to the user [9, 10 and 11]. Several types of search engines available are crawler based search engines, human powered directories, meta search engines, hybrid search engines and domain specific search engines.

2.1 Domain Specific Web Search Engine
A search engine that runs over all domains must give equal weight to all words. However, if a search engine is domain specific it is intuitive that some terms are more important than others. The downside to most of these domain specific search engine is that they tend to use syntax-based mechanisms and most of information published on web is in HTML file format. On the bases of these mechanisms most of the results are irrelevant and not accurate to users query.

2.2 Ontology
The term ontology is a data model that represents a set of concepts within a search engine and the relationships between those concepts. Ontology is a formal description of concepts and the relationships between them. Web Ontology Language (OWL) and Resource Description Framework (RDF) is used for construction of ontology [8].

An ontology is an explicit specification of a conceptualization and a systematic account of existence, what "exists" is that which can be represented. An ontology is the statement of a logical theory. A commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology [3].

Ontology can be used to define common vocabularies for specific domain and to share common understanding of the structure of information. It enables reuse of domain knowledge, separate domain knowledge from operational knowledge and analyze domain knowledge [4]. Ontology is mainly domain specific as: education, medical etc.

2.3 Semantic Lexicon
Semantic similarity is the likeness of the semantic content among documents or the meaning among a set of structured terms. The semantic relations into which a word enters determine the definition of that word. A sematic lexicon is a dictionary of words labeled with semantic classes and it create a set of sematic relation between terms called synsets [13, 14].

Synonymy is WordNet basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses. Synonymy ("syn" - same, "onyma" - name) is a symmetric relation between word forms [5, 7, 17, and 15].

2.4 Ontology Based Search Engine
A domain based search engine uses ontology to enhance its results. This design has an ontology that describes the area of

interests of search engine. These ontology uses weight concept, in order to prioritize some words over other words and relevance score that separate relevant pages from irrelevant page. Which on combine overcomes the drawbacks of search engine and results returned are more accurate.

Ontology enhances the vocabularies of search engine and results to the user's quires. Ontology based search engine will be able to minimize the chance of producing irrelevant result.

3. PROPOSED APPROACH

The concepts of ontology act as an intelligent filter for search engines. Search engines are web service runs on the mechanism of syntax based. Ontology acts as a service that enhances the traditional search mechanism by implementing the concepts of ontology.

Traditional domain specific search engines that make use of syntax or keyword-based search fails to provide accurate search results because of irrelevant pages are being indexed with relevant pages in database. Ontology based domain specific search engine can retrieve accurate and relevant information. These concepts are also used to improve the quality of relevant search results by ranking the relevant web links on their relevance score.

The main aim of this paper is to describe designing process of domain based search engine using ontology. This section explains each step and shows the process of creating ontology (ontologies, syntable, weight table and calculation of relevance score). Figure 1 shows the low level diagram of domain specifc search engine using ontology.

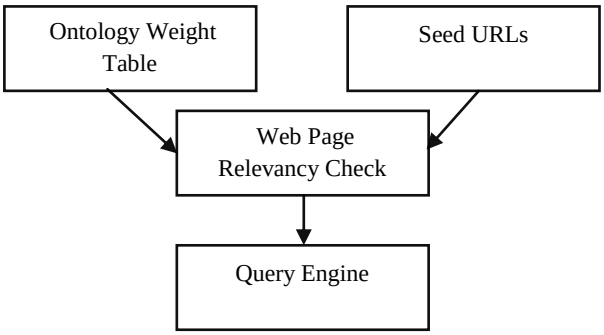

Figure 1. Low Level Diagram Of Domain Specific Search Engine using ontolgy

3.1 Domain Ontology Terms

To separate domain knowledge from operational knowledge and for analyze of domain knowledge it uses OWL. OWL is a family of knowledge representation languages for authoring ontologies. OWL ontologies can import other ontologies, adding information from the imported ontology to the current ontology. OWL describes classes, properties and relations of the structure of information among people or software agents for share common understanding [8, 16].

OWL resembles class hierarchy form. Class hierarchies are meant to represent structures, re-use of data, mix of data. Ontologies are structured in a hierarchical form which serves as syntable and weight table database. Vegetable domain ontology using hierarchical form is shown in Figure 2.

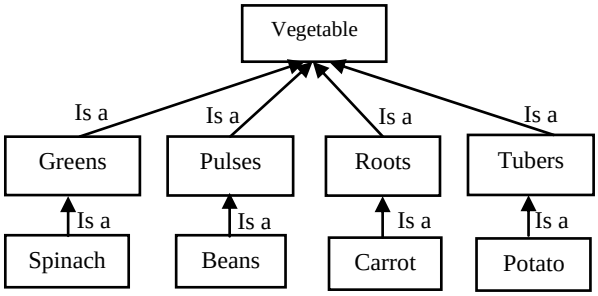

Figure 2. Vegetable Domain Ontology in Hierarchical Form

3.2 Syntable

Syntable **(St)** is a table that contain the *synonyms term* **(X**s) of the ontology term. Syntable provides a common understanding of a term and also its relationship with other terms (i.e. easy to validate) and set of synonyms called synsets.

All synsets are connected to other synsets by means of sematic relation. A sematic lexicon is a dictionary of words labeled with semantic classes all come under WordNet. WordNet into a lexical ontology usable for knowledge representation. Knowledge representation is a tool that accurately uses a set of symbols to represent a set of facts within knowledge base. Knowledge can be represented using semantic net. Semantic data is human understandable as well as machine understandable which makes it more efficient standard for data representation.

Semantics is the study of meanings of concepts, terms or words. Syntactically different terms can be used to represent the same concept using ontology [6]. Syntable for vegetable domain is shown in Table 1 (Forward slash "/" shows empty or end of list).

Table 1. Syntable (St (1))

Ontology terms(X)	Synterms (Xs)
Potato	Aalloo
Carrot	Gaajar
Bean	/
/	/

3.3 Weight Table

Weight table **(Wt)** separates specific term from common term, prioritize specific term over common term and are used for relevance score calculation for any web page content.

Ontologies are being used to construct weight tables. It has two columns, first column is for *ontology term* **(X)** and second columns is for *term corresponding weights* **(W)**. Weight will be assigned between "0" to "1". The strategy of assigning weights [2,12] depend on the basic idea that the more specific term to domain is assigned with more weight and other terms which are common to more than one domains are assigned with less weight as described in Table 2 ("/" shows the end of table).

Table 2. Weight Table (wt(1))

Ontology term(X)	Weight value (W)
Potato	1.0
Carrot	0.8

Beans	0.4
/	/

3.3.1 Weight table function

The flowchart of weight table function is shown in Figure 3.

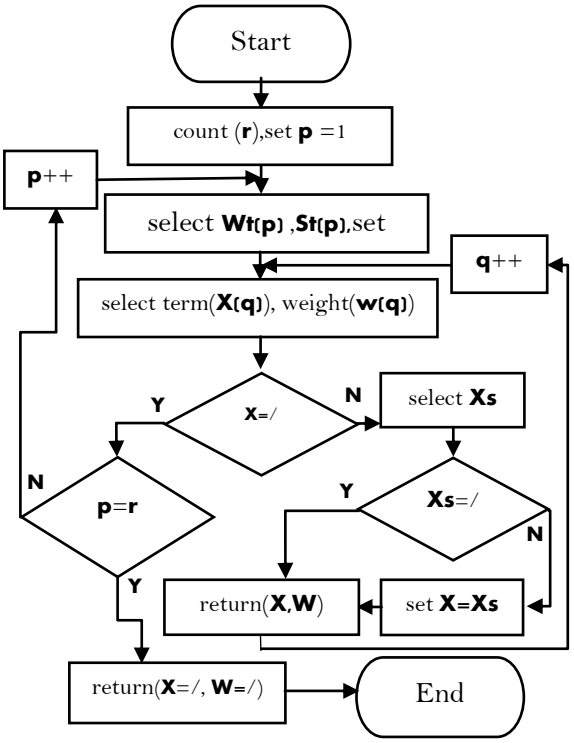

Figure 3. Weight Table Function Flowchart

This function returns term (X) and corresponding weight value (w). Weight table function runs until all table traversed and finally return "/" (which means end of ontology terms). Here **"r"** is the number of weight tables.

3.4 Relevance Score

Relevance score is a numeric value for each webpage, which is generated on the basis of the term weights value, term synonyms, number of occurrence of ontology terms. It is used for consideration of web page to specific domain or not. Relevance score can also be used to check rank of web links.

Relevance score **(Rs)** of a web page can be calculated by using web page relevancy check function as shown in flowchart Figure 4.

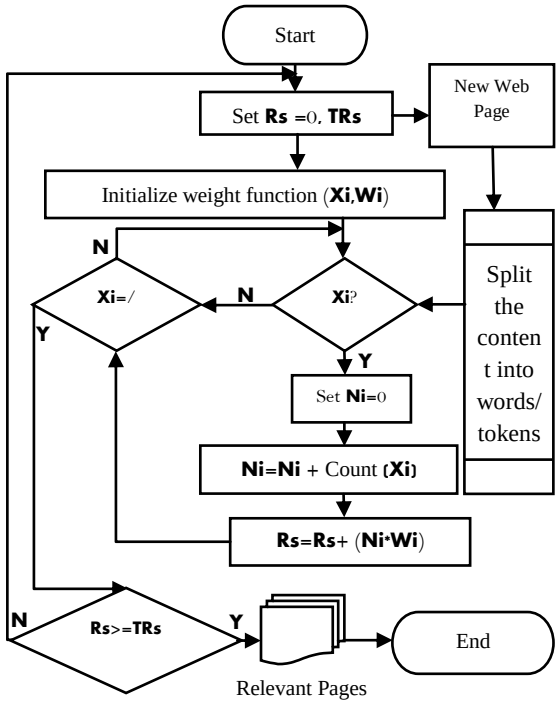

Figure 4. Web Page Relevancy Check Function.

After calculating the relevancy of web page, compare it with the *threshold relevance score* **(TRs)** for consideration of web page to specific domain.

Now, to find *relevance score* **(Rs)**, it first call weight table function. After calling the weight table function it *count* number of *occurrence* **(Ni)** *of term* **(xi)**. Then multiply counted number of occurrence with *corresponding weight value* **(Wi)** and saves in *relevance score* (Rs), **Rs = Rs + (Ni*Wi)**. Finally comapare the relevance score (Rs) with threshold relevance score (TRs) in order to consider page, **Rs >= TRs**.

3.5 Detailed Architecture

The detailed architecture of proposed domain specific search engine using ontology is shown in Figure 5.

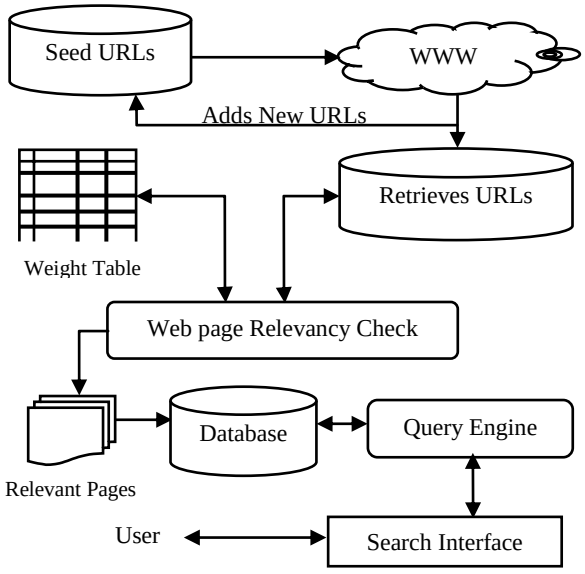

Figure 5. Detailed Architecture

This architecture emphasizes on calculating relevance score and come out with relevant pages. This system adds calculation of relevance score of web pages and all web pages that have relevance score greater than or equal to a specified threshold relevance score are consider relevant for specific domain. These web links are then saved in search engine database.

Seed URLs are serves as input for crawler and performs crawling function. The crawling function use tolerance limit. Tolerance limit computed the depth limit of web links to crawl. After the completion of crawling function it produces web links that are relevant for specific domain. Now, in relevance function, relevance score is calculated for all related web links. Than these relevance scores are compared with threshold relevance score.

When user makes a request, the request is send to the query engine. The query engine then uses page selection attributes and check the efficiency of the selected pages. The page selection attributes are: pages that contain only search term, pages which include only query form (exclude non-query), and all redundant query form within same pages are removed. Finally the appropriate web links get selected and results are displayed.

4. CONCLUSION AND FUTURE WORK

Syntax based search methods are not enough to produces relevant results. Ontology makes common understanding and structure of web contents. In this paper architecture of a domain specific search engine using ontology has been presented. Use of ontology returns better results. Weight table and relevance score function are used to calculate relevance score and rank. In future improved algorithm can be used for updating weight table and ranking pages. This work can also be extended for multiple domains.

5. REFERENCES

[1] T. R. Gruber, "A translation approach to portable ontologies", Knowledge Acquisition, Vol. 5, Issue. 2, pp. 199-220, 1993.

[2] D. Mukhopadhyay and S. Sinha, "A New Approach to Design Graph Based Search Engine for Multiple Domains Using Different Ontologies", pp. 1-8.

[3] T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal Human-Computer Studies, Vol. 43, pp. 907-928, 1993.

[4] N. F. Noy and D. L. McGuiness, "Ontology Development 101: A Guide to Creation Your First ontology", 2015.

[5] WordNet, www.wikipedia.org/wiki/wordNet; 2015.

[6] J. M. Rajan and M. Deepa Lakshmi, "Ontology based Semantic Search Engine for Healthcare Services", International Journal on Computer Science and Engineering, Vol. 4, No. 04, pp. 589-594, 2012.

[7] A.Upadhyay, A. Paul and Pijush K. Dutta Pramanik, "Semantic Web Crawler for More Relevant Search Using Ontology", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue. 12, pp. 958-962, 2014.

[8] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/; 2015.

[9] Rashmi Agarwal, P. sagar and Niraj Singhal, "Performance Metrics for Selection of Quality Hidden Web Documents", International Journal of Advanced Research in Computer Science, Vol. 5, No. 7, pp. 248-252, 2014.

[10] Niraj Singhal and Ashutosh Dixit, "Web Crawling Techniques: A Review", Proceedings of Native Seminar on Information Security: Emerging Threats and Innovations 21st Century, India, pp. 34-43, April 2009.

[11] A.Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan, "Searching the web", Journal of ACM Transactions on Internet Technology, Vol. 1, Issue. 1, pp. 2-43, 2001.

[12] D. Mukhopadhyay, A. Biswas and S. Sinha, "A New Approach to Design Domain Specific Ontology Based Web Crawler", proceedings of IEEE 10th International Conference on Information Technology, pp. 289-291, 2007.

[13] Vinit Kumar, Niraj Singhal, Ashutosh Dixit and A. K. Sharma, "A Novel Architecture of Perception Oriented Web Search Engine based on Decision Theory", Indian Journal of Science and Technology, Vol. 8, Issue. 7, pp. 635–641, 2015.

[14] Noah SA, Che AA and Zakaria LQ, "A semantic retrieval of web documents using domain ontology", International Journal of Web and Grid Services, Vol. 1, Issue. 2, pp. 151–164, 2005.

[15] G. A. Miller, "WordNet: A Lexical Database for English", Communication of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[16] G. Antoniou and F. V. Harmelen, "Web Ontology Language: OWL", pp. 1-26.

[17] A. Soni, H. Sunhare, S. Patel, "Semantic Retrieval Technique Based On Domain Ontology", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 7, pp. 3187-3192, 2013.