The Use of Ontology in Semantic Search Techniques

Prerna Parmeshwaran
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Juilee Rege
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Sindhu Nair
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

ABSTRACT

The World Wide Web has grown over the years from simple hypertext documents to highly interactive pages, where users can also contribute to the content by posting comments and so on. However, most data is extremely unstructured and cannot be easily automatically processed by machines. Presently, most search engines are keyword based and searches may also result in irrelevant results due to the mere presence of matching keywords. To eradicate this problem, the concept of semantic web has been introduced in which the data follows a uniform standard. Everything present in the document has a specific meaning attached to it. Such standardized documents can easily be understood by machines. Due to the concept of semantic web, search engines can be made to understand the meaning of the query and thus the most relevant links can be retrieved. To implement semantic web technologies, the concept of ontology is used. In this paper, an attempt is made to explore how semantic web and ontology are being used to implement efficient search engines.

General Terms

Semantic web, searching, page ranking

Keywords

Ontology, search engines, semantic web, semantic annotation, semantic indexing

1. INTRODUCTION

The World Wide Web is a huge repository of hundreds of thousands of articles and information on almost any topic. It seems it would be easy to find answers to any question one has by just entering the query into the search bar and, through complex algorithms designed to find the best solutions, hundreds of replies are shown on the screen in a matter of milliseconds. But this is not true. The information is scattered, unstructured and even inconsistent. The most basic search to find information is a keyword search in which the search tool lifts words from the query and matches them to words in documents in its repository. As such, the articles with the maximum number of words matching the query are chosen and shown. This can be highly inefficient as it can result in a lot of unwanted material. For example, the query ‘books written by Bill Gates’ is a very specific query requesting only those books which have been authored by him. But since the searching tool picks out keywords, ‘books’, ‘written’, ‘Bill Gates’ would be picked out and articles even containing information on books written on him would be shown to us. As such, it could also happen that the information the user needs might go lower on the pages or even on the next pages. This is because the searching tool picks keywords and not the actual concept of the query. Relationships between the words are not considered and so written by Bill Gates is ignored.

To solve this limitation of keyword search, the idea of semantic search has been introduced. Using the concepts of semantic web, highly efficient search engines that retrieve only the most relevant results can be built. Such search engines will also facilitate the implementation of extremely intelligent expert systems which base their decisions on the results of a query.

2. METHODOLOGY

2.1 Ontology

Ontology is the technique of defining names, attributes and relationships between items of a particular domain. Domain ontology is the term used to describe the rules and constraints in a particular field and helps to find the associations between terms. Thus ontology can effectively be used to assign meanings to words in the semantic web.

There are several tools available for the construction of domain ontology. Protégé, a java-based open source ontology editor is now widely used for domain ontology construction. Protégé supports the Web Ontology Language (OWL) which is popularly used to represent entities and their relationships. OWL is extremely easy to use and it is more enhanced than other formats like Resource Description Framework (RDF) and Extensible Markup Language (XML).

2.2 Architecture of a semantic search engine

The general architecture of a semantics based web search engine that has been proposed [1] is as follows:

- Crawler: The crawler collects the web pages from different domains. A web database is constructed to store these pages for future retrieval.
- Pre-processor: The web pages returned by the crawler are usually unstructured and need to be pre-processed before the construction of ontology. Meaningless words are removed and only the most relevant words are retained. Also, HTML tags of no importance are removed.
- Semantic annotator: This module aims to generate metadata in order to assign actual meanings to the document and its entities. By formally describing these entities, the web search will be associated with the meanings of words and not just pattern matching with the given keywords. Annotation graphs are constructed where the concepts are converted to graph nodes and the relations to edges. The nodes are joined by an edge only if the concepts are related to each other. Thus, many relevant classes in a particular domain are defined and characterized by connected concepts.

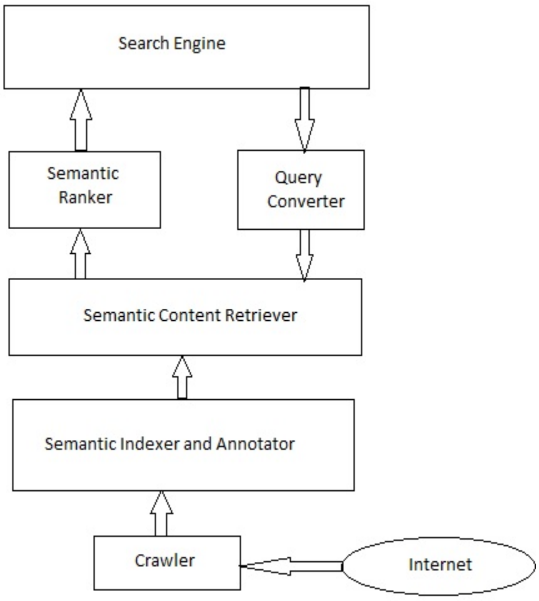

Figure 1. Architecture of a semantic search engine [1]

- Semantic indexer: The web documents along with the metadata are now indexed with the semantic entities. To check how well a web document is mapped to the concept behind the entity, a mapping score is computed. This mapping score is a function of the frequency of keywords, frequency of concept and also the frequency of keywords and concepts with HTML tags such as metadata, anchor and so on. Two techniques are used [3] to perform this task: term weighting and similarity measure.

Term weighting: It calculates the relevance or importance of a word in a document. It uses Term Frequency (TF) to calculate frequency of the word in the document (local weighting) and Inverse Document Frequency (IDF) to calculate the relevance of a word n the entire collection of documents. TF*IDF provides a good measure of the importance value of a word in a pool of related documents. The standard formula used [3] is:

$$TF_{Di} = \frac{\sum occ(w)}{card_{Di}} \qquad (1)$$

Similarity measure: Two types are calculated: The distance measured in the vectorial space and the cosine measure which calculates the similarity between the query provided and the documents available. These mapping scores are used to rank pages according to relevance at a later stage.

The formula for computing the distance in vectorial space [3] is:

$$Dist(Q_k, D_j) = \sum_{i=1}^{T} | q_{ki} - d_{ji} | \qquad (2)$$

The formula for the cosine measure [3] is given by:

$$RSV(Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki} d_{ji}}{\sqrt{\sum q_{ki}^2 \sum_{i=1}^{T} d_{ji}^2}} \qquad (3)$$

- Semantic query convertor: Here the query is expanded and the search happens in three ways: by concept, by links, and also by searching for keyword with similar meanings (thesaurus).
- Semantic content retriever: In this module, the appropriate results are extracted from the semantically indexed web content obtained earlier. The retrieved content is matched by keyword as well as the concept, and the final result is the intersection of the documents containing the keywords and the concepts.
- Semantic ranker: The ranking of results must be done so that the user is able to view the most relevant results first. The documents that are most similar to the query are ranked higher.

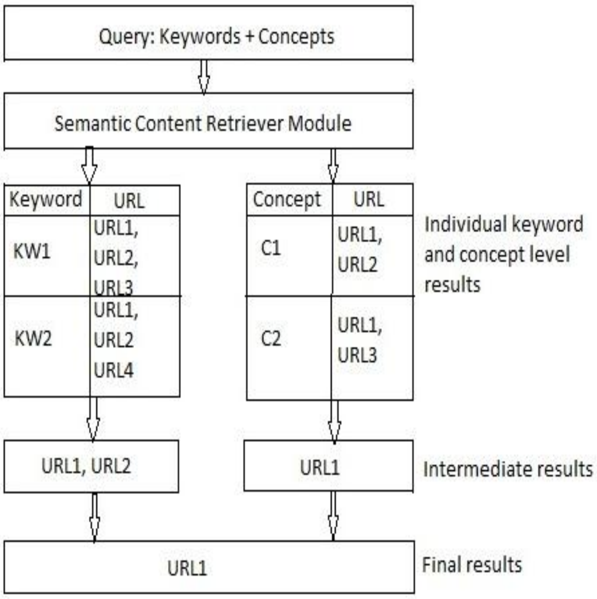

Figure 2. Semantic content retriever [1]

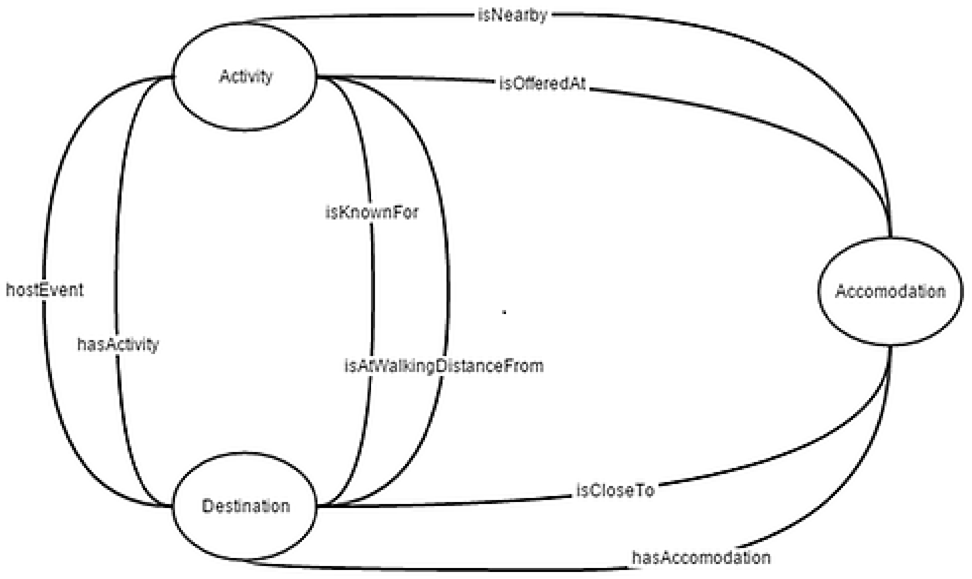

Figure 3. Example of an annotated graph [4]

2.3 Page ranking techniques

The ranking of the pages can be done with the aid of the annotated graphs as shown in figure 3. The strategy presented [4] is shown below:

The annotated graph of a domain is known as a Query Graph G and is defined as G(C, R) where C is the set of concepts i.e. total number of nodes in the graph and R= (Rij|i=1…n, j=1|….n, j>i) is the set of relations between two concepts i and j i.e. the edges between nodes i and j. Given the query by the user, a subgraph GQ can be created. Both the graph G and subgraph GQ are used in ranking the relevant pages.

For example: Consider 3 keywords, k1, k2, k3 and their relevant concepts c1, c2, c3. Suppose there are two pages P1 and P2 which contain all three of these sets. Thus these pages have to be ranked. If the graph depicts that for the first page, c1 and c2 are linked with c3 through one relation and for the second page, there are two relations between c1 and c3 and no link between c1 and c2. The probability between c1 and c2 and c1 and c3 is calculated.

According to the probability theory, P(rij,p)= sij/nij=Tij. Thus the relational probability

For the first page,

P(r13,p1)=s12/n12=T12=1/2,

P(r13,p1)=s12/n13=T13=1/2.

For the second page,

P(r1,p2)=s12/n12=T12=0,

P(r13,p2)=s13/n13=T13=1.

Compute the joint probability P (Q,p)=p((r12,p)n(r13,p)),

For the first page is P(Q,p1)=¼.

For the second page is P(Q,p2)=0.

Thus the pages are ranked and first page is placed before second in the result.

Another page ranking algorithm known as the top-k algorithm [6] recognizes the fact that typical users of the Web are only interested in the top k queries returned by the search engine. This algorithm is also based on ontology and semantic relationships. To determine the relevance of a particular result, the following measures are used:

- Number of meaningful semantic paths: A resource that is semantically matching with the keywords could be more important. Higher weights are assigned to the semantic paths that are directly related to those paths that are directly linked to the user's query.
- Number of keywords covered: A resource that is connected to as many keywords in the user's query through semantic paths is considered to be important.
- Discriminating power of keywords: A resource that has semantic paths to the query and is noticeably different from the other results is also considered to be relevant.

Using these three relevance criteria, the retrieved results are ranked and the top k results are ultimately displayed to the user. It has been recognized that while this method is effective in reducing the search space, it is not always practical to scan a huge instance graph for each query. To overcome this problem, an offline pre-processing method [6] involving a keyword index is used. This index maps keywords to relevant resources. An index list for a keyword contains the type, identifier and a relevance score. Each list is sorted according to the descending order of the relevance score. If there are m keywords in the user's query, m index lists (each list corresponding to the keywords in the query) are analyzed. Then, the top k results with the highest overall relevance score are displayed.

3. CONCLUSION

The repository of data available to us over the World Wide Web is now termed as big data due to its sheer vastness. There are thousands of links and results for a single query and it is very important that only the most relevant results are presented to the user. Semantic based searching helps to increase the efficiency of relevant page gathering, analyzing, ranking and ultimately displaying, in the decreasing order of importance and closeness to the submitted query, the final pages to the user. This kind of search can be effectively used in creating expert systems used for interpretation, prediction, design, planning, or even repair and control. Future plans can include implementation of the techniques on real world problems such as medical diagnosis expert system or agricultural production expert system.

4. ACKNOWLEDGMENTS

The authors would like to acknowledge Dwarkadas J. Sanghvi College of Engineering and the Computer Engineering Department for their support and co-operation.

5. REFERENCES

[1] M. Thangaraj and G. Sujatha, 2014, An Architectural Design for Effective Information Retrieval in Semantic Web, In Expert Systems with Applications, Volume 41, Issue 18.

[2] Ziang Li, Wei Xu, Likuan Zhang, Raymond Y.K. Lau, 2014, An Ontology based Web Mining Method for Unemployment Rate Prediction, In Decision Support Systems, Volume 66.

[3] Abdelkrim Bouramoul, Mohamed-Khireddine Kholladi, Bich-Lien Doan, 2012, An ontology-based approach for semantics ranking of the web search engines results, In 2012 International Conference on Multimedia Computing and Systems (ICMCS)

[4] J.Anitha Josephine, S.Sathiyadevi, 2011, Ontology Based Relevance Criteria for Semantic Web Search Engine, In IEEE.

[5] Ian Horrocks, 2008, Ontologies and the Semantic Web, In Communications of the ACM, 51(12):58-67.

[6] Jihyun Lee, Jun-Ki Min, Alice Oh, Chin-Wang Chun, 2014, Effective ranking and search techniques for Web resources considering semantic relationships, In Information Processing and Management, 50.

[7] Damaris Fuentes-Lorenzo, Norberto Fernandez, Jesus A. Fisteus, Luis Sanchez, 2013, Improving Large Scale Search Engines with Semantic Annotations, In Expert System with Application, Volume 40, Issue 6.

[8] Khadija M. Elbedweihy, Stuart N. Wrigley, Paul Clough, Fabio Ciravegna, 2015, An overview of Semantic Search Evaluation Initiatives, In Web Semantics: Science, Services and Agents on the World Wide Web, Volume 30.

[9] Cai Bo, Li Yang-Mei, 2014, Design and Development of Semantic-based Search Engine Model, In International Conference on Intelligent Computation Technology and Automation (ICICTA).

[10] Alexandros Batzios, Pericles A. Mitkas, 2012, WEBOWL: A Semantic Web Search Engine Development Experiment, In Expert Systems with Applications, Volume 39, Issue 5.