Asian Journal of Multidisciplinary Research (AJMR)

E-ISSN: 2395-1702 P-ISSN: 2395-0382
Volume 01- Issue 01-, pp-18-26

Research Paper Open Access

THE SEMANTIC SEARCH ON WEB USING ONTOLOGICAL RELATIONSHIPS WITH RANKING THE DOCUMENTS

S.Gopalakrishnan [1], Mr.S.Divakar [2]

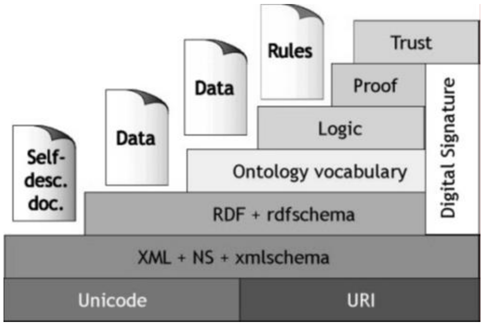
M.E., Department of Computer Science, Arunai Engineering College,
Tiruvannamalai, India
Assistant Professor, Arunai Engineering College, Thiruvannamalai, India
gkcsresearch@gmail.com,reachdivakar@gmail.com

Abstract— the search engines have become the most powerful tools for obtaining useful information scattered on the web and the current web Information Retrieval system retrieves relevant information only based on the keywords which is inadequate for that vast amount of data. The research on semantic search aims to improve traditional user query and retrieval of relevant information. The Semantic Web is an evolving development of the World Wide Web in which the meaning (semantics) of information and services on the web is defined, making it possible for the web to "understand" and satisfy the requests of people and machines to use the web content. The Elements of the semantic web are expressed in formal specifications that include Resource Description Framework (RDF), a variety of data interchange formats e.g. RDF/XML and notations (RDFS) and the Web Ontology Language (OWL) are intended to provide a formal description of concepts and relationships.The architecture takes as input a plain of keywords by the user and query is converted into semantic query with the help of domain ontology and discovers semantic relationships between the runtime. The relevant information for the semantic query is retrieved and ranked according to improved Semantic Annotation and Indexing algorithm. The performance analysis shows the proposed system can improve the accuracy and effectiveness for retrieving relevant web documents compared to the existing systems.

Keywords — Information retrieval, Semantic search, semantic query, ontology;

1. INTRODUCTION

The World Wide Web has changed the way people communicate with each other and the way business is conducted. It lies at the heart of a revolution which is currently transforming the developed world towards a knowledge economy, and more broadly speaking, to a knowledge society. This development has also changed the way we think of computers. Originally they were used for computing numerical calculations. Currently their predominant use is information processing, typical applications being data bases, text processing, and games. Most of today’s Web content is suitable for human consumption. Even Web content that is generated automatically from data bases is usually presented without the original structural information found in data bases. Typical uses of the Web today involve humans seeking and consuming information, searching and getting in touch with other humans, reviewing the catalogue of online stores and ordering products by filling out forms, and viewing adult material. Apart from the existence of links which establish connections between documents, the main valuable, indeed indispensable, kind of tools are search engines.



Keyword-based search engines, such as AltaVista, Yahoo and Google, are the main tool for using today's Web. It is clear that the Web would not have been the huge success it was, were it not for search engines. However there are serious problems associated with their use.

Ontology is an explicit and abstract modeled representation of already defined finite sets of terms and concepts, knowledge engineering and intelligent information integration. Ontology defines "explicit section of conceptualization and the concept, relationships, and other distinctions that are relevant for modeling a domain. The Specification takes the form of the definitions of representation vocabulary (classes, relation, and identity), which provide meanings for the vocabulary and formal constraints on its coherent use.

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats.

The RDF data model is similar to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions. These expressions are known as *triples* in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

The RDF between the <rdf:Description> tags is called an RDF statements.

- Subject
- Object
- Predicate of the statement

The namespace "//www.w3.org/1999/02-rdf-syntax#" that we find, is standard w3.org namespace. For example, one way to represent the notion "The sky has the color blue" in RDF is as the triple: a subject denoting "the sky", a predicate denoting "has the color", and an object denoting "blue". RDF is an abstract model with several serialization formats (i.e., file formats), and so the particular way in which a resource or triple is encoded varies from format to format.

2. Related work

The unsolved problems of current search engines have led to the development of semantic web search system Conceptual search has been the motivation of a large body of research in the IR field long before the semantic web vision emerged. ''Semantic search'' is a layered architecture that separates end users from the back-end heterogeneous semantic data repositories. ''Semantic search'' accepts keywords as input and delivers results which are closely relevant to the user keywords in terms of

Parameter	Traditional Keywords search	Semantic based search
Dataset	Documents	RDF triples, semantically annotated documents
Data organization	Unstructured	Semi- structured
Search orientation	Document centric	Entity, relationship and semantic document centric
Collection	Bag of words	Collection Bag of words
Query processing approach	Matching and filtering	Not just matching and filtering but also joining
Domain of satisfaction	Work well for topical search	Complex queries are satisfied, more precise answers
Scalability	Web scale	Not scale to massive and heterogeneous Web environment

semantic relations. The Semantic search with a ranking algorithm designed specifically for an ontology-based information retrieval model with a semantic indexing structure based on annotation weighing techniques. The inherited relationships between the keywords are analyzed in terms of concepts in from these concepts and relations a concept-relation graph is formed which is used to eliminate the less ranked arcs. It also creates a property-keyword candidate set and sent it to the web page database to get a retrieved result set for the users. The efficiency of this approach is limited by lack of ranking technology. This motivates a relation based page ranking algorithm for semantic web search. The ranking technology is based on the estimate of the probability that keywords/concepts within an annotated page are linked one with another in a way that is the same to the one in the user's mind at the time of submitting the query. The probability is measured using a graph based description of ontology, user query and the annotated page. In these approaches further efforts are requested for future semantic web repositories based on multiple ontology's and better ranking. By building upon a dynamic ontology our model supports multiple domains with semantic dynamic ranking.

The cluster based approach for information retrieval provides features in terms of reduced size of information provided to the end users. The clusters of items with common semantic and/or other characteristics can guide users in refining their original queries. Users can zoom in on smaller clusters, and then drill down through subgroups. Whereas this work is concerned with the query expansion, SBIRS is concerned with starting from query expansion to retrieve ranked results.

A Crawler-based indexing and information retrieval system for the semantic web Swoogle extracts meta data for each discovered document, and computes relations between documents. The ontology rank is computed as a measure of the importance of a semantic web Document. Swoogle is improved by adding user preferences and interests to provide user a set of personalized results. Swoogle is strictly for semantic web documents whereas semantic web approach converts web documents into semantic web documents.

A search engine that uses several mapped RDF ontologies for concept based text indexing is discussed in for any information retrieval system ranking algorithm is defined with certain metrics. The variety of relevance ranking metrics are discussed and analyzed and It proposes a set of metrics to estimate the personal, topical and situational relevance dimensions. These metrics are calculated mainly from contextual information and usage and do not require any explicit information from users. Our work move from the consideration above and relies on the assumption that for providing effective ranking the semantic web is logic makes use of the underlying ontology and of the web page to be ranked in order to compute the corresponding relevance score.

A semantic-based approach to content annotation and abstraction for content management is proposed and In this approach a semantic-driven content environment which features a high interoperability of content can be constructed for bridging the semantic gaps for the customer and the content author to increase the efficiency of content management. This work is improved based on the semantics consisting of elements like subject, predicate, and object. In this work a semantic pattern expression capable of representing content semantic features has been designed to represent human semantics with topic-to-topic associations and to replace keywords as the input of the information retrieval system. This architecture contains the core technologies such as Semantic determination and extraction, Semantic Extension, Semantic Pattern and matching. Compared to these approaches, the proposed architecture considers web documents and a much more detailed, densely populated conceptual space in the form of ontology based knowledge base and thesaurus instead of a topic map.

The popular ranking model for the ordering of retrieved documents is PageRank(Page, Brin, &Motowani, 1998). It looks at the internet as a big graph where pages are nodes and hyperlinks are edges. It has been applied to distinguish the popularity of different web document through analyzing the links structure in the web graph. The web pages involving same keywords are not equally popular.

In order to help user to quickly locate their pages of interest, popularity of retrieved pages is required to be calculated. The more popular a web page is, the more likely the user is interested in the more important that pages and PageRank algorithm facilitates to accurately such global importance for given page. It is based on the intuition that more the number of random visitor and reference to a page are, the more popular of the page process, query context is taken into ranking of the page to given user query.

3. Proposed semantic web architecture

The architecture for information retrieval from semantic uses conceptual representations of content beyond plain keywords as knowledge bases and provides conceptual representations of user needs. This architecture handles the concept representations of the content, query extensions, matching the semantics, extraction of the relevant results in the order of relevancy with the help of the following components.

- Crawler
- Preprocessor
- Semantic annotator
- Semantic indexer
- Semantic query converter
- Semantic content retriever
- Semantic ranker

These components are grouped under different layers of the architecture. To creates web database with the components crawler, preprocessor.

The Semantic Annotation that creates knowledge base with semantic annotator and indexer. Semantic Matching are performs matching between semantic content and the semantic query. The retrieval processes are ranks the retrieved results and is submitted to the user application. The overall architecture with the above said components are given

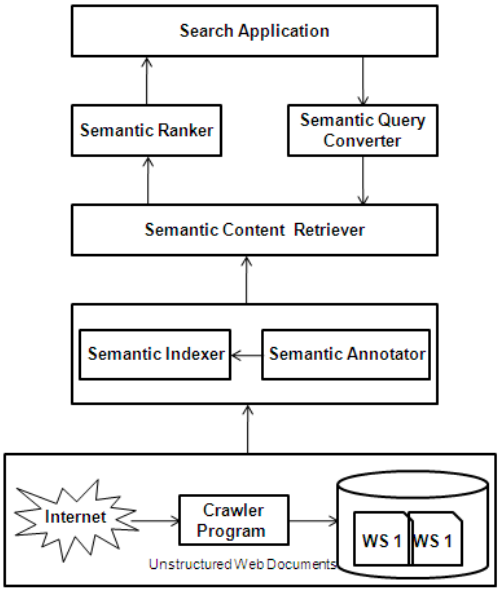


Fig2. Semantic Web Search Architecture

3.1 Web crawler

The crawler collects the web pages from different domains. The collected web pages are stored to a web database for the use of retrieving URLs and corresponding web pages. The result of the crawler is sent to the preprocessor for getting the pure content from this unstructured web documents.

Web Crawlers roam the web sites with the aim of automating specific tasks related to the web and they are responsible for collecting the web-content.

3.2 Pre-processor

The unstructured web documents are pre-processed before matching into ontology concepts and Using HTML parser the meaningless HTML tags are removed. After extracting text from the web documents the less meaningful words known as stop words like neuter pronouns, articles, and symbols are removed.

Stemming is the process for reducing inflected (or sometimes derived) words to their system, base or root form generally a written word form.

Unstructured data and Semi-structured data and structured data:

		Id	Name	Age	Degree
The University has 5600 students. John ID is number1; he is 18 years old and already holds a B.Sc., degree. David ID is number 2, he is 31 years old and holds a Ph.D., degree. Robert ID is Number 3, he is 51 years old and also holds the same Degree as David,a Ph.D degree.	<University> <Students ID=1> <Name>John</name> <Age>18</Age> <Degree>B.Sc </Degree> </Student> <Student ID=”2”> <Name>David</Name> <Age>31</Age> <Degree>Ph.d<Degree> </Student> </university>	1	John	18	B.Sc.
		2	David	31	Ph.D.
		3	Krish	28	M.E.
		4	Robert	51	Ph.D.
		5	Michal	21	B.E

While processing documents, this preprocessor will filter images, audio, video and other information formats, and will identify and eliminate the noise content. The same process is repeated for the user query on stop words and derived words.

3.3 Semantic annotation

Its type of Meta data generation and usage schema used to extend the existing information access methods. The annotation scheme used here is based on the concepts of the particular predefined domain ontology and the meanings of the phrases as semantic entities.

Those entities can be coupled with formal descriptions and thus provide more semantics and connectivity to the web database with the help of the domain ontology's,

3.4 Semantic indexer

The annotated web documents of the knowledge base resulted from semantic annotator are indexed with the semantic entities. The mapping score which indicates how good a web document is mapped to an ontological concept is computed. And hence the indexer creates a weighted semantic annotation/indexing

Scores are computed by an adaptation of improved TF-IDF algorithm. The weight is a function of term frequency of the keyword [tf (W)], term frequency of concept [tf (C)], Tag based keyword frequency [tagf (W)], and Tag based concept frequency [tagf (C) and normalization factors. For the tag based frequency specific tags of the HTML file such as Head, Title, Meta, Description, Anchor tags are considered. The importance will be given for the presence of keyword and/or concept in the URL of the web page. The weights thus calculated are used for ranking.

3.5 Semantic query converter

The plain keywords entered by the user is expanded and converted into semantic query in three different ways.

(i) By matching the concepts in the domain ontology.

(ii) By using the links of the websites with the help of an automatic link extraction algorithm.

(iii) By using the thesaurus. The Extended semantic query is presented to the user as ontology suggestion.

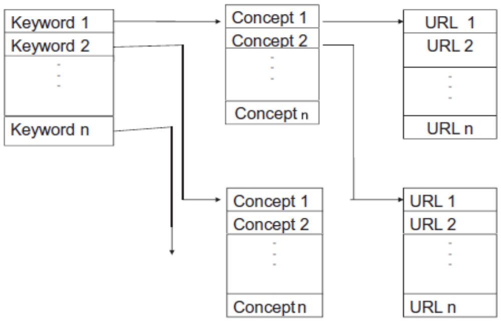


Fig 3. Mapping of keywords with concepts of ontology

The user by making use of the suggestion selects his concept of searching and submits it to the semantic content retriever. The user query extension with the concepts and matching of web documents.

3.6 Semantic content retriever

This component concerns with identifying and submitting the most approximate content to query by matching in the semantic content for the query semantic patterns, and it covers the following steps: The semantic query from the semantic query converter is matched here with the semantically indexed web content. The retrieved content should be matched with the two parts (keyword and concept) of the semantic query.

The final retrieval list is the intersection between the set of the web documents containing the keyword and the semantic entity/ contextual meaning. To retrieve the intersecting list a hash table structure which is having two columns is used.

The first column has the Query words of the extended query and the second column has the list of web documents that matches with that Query words. This process is depicted in the given fig. The resultant list will be the intersection of the web documents that are stored in the second column.

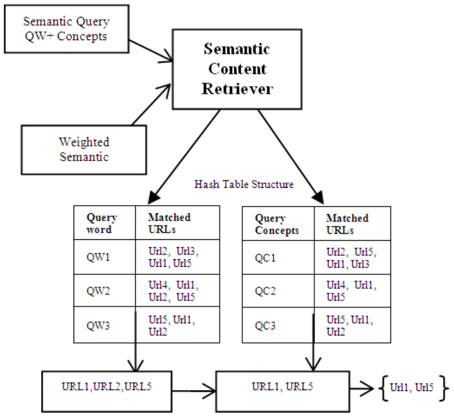


Fig 4. Semantic Content Retriever

3.7 Semantic ranking the documents

The retrieved list of the previous module is ranked with the help of the semantic weights. The relevancy is measured with the weights calculated by the improved dynamic ranking algorithm. The weight is a function of term frequency, collection frequency and also some normalization factors. The term frequency is the local weighting factor which reflects the importance of the term within a particular document.

The global weighting factor considers the importance of a term within the entire collection of documents known as document frequency (df). Next the inverse document frequency (idf) which relates the document frequency to the total number of documents in the collection (N) is computed.

In contrast with the keyword based system these values will be calculated for the keyword as well as for the semantic entities. Now the similarity coefficient (sc) between the query and the web document is defined by the dot product of weights of the corresponding words and semantic entities of the semantic query (n) and the web document.

Based on the algorithm the weight depends on the two main factors. One is the quoted frequency of the keywords in the web documents and another one is the semantic entities in the content of the web pages. When there is no semantic entity the retrieval is nothing but the keyword based.

The improved algorithm based on TF-IDF algorithm can fit both traditional web and semantic web making the IR more accurate and promote the efficiency and the precision of traditional web search and semantic web search.

4. Algorithms

Algorithm: Semantic Annotation and Indexing

Input: Set of Web Documents (D) in a particular domain (N): Set of concepts from domain ontology

Output: Semantic Content Knowledge Base from user queries with its score

Parameters: N-Total Number of Web documents,

D-Set of Web documents, S-Stop words, wij – jth word in ith document, m-Number of Keywords in a Web document,

tf-Term Frequency, tagf-Frequency of terms in HTML tags,

c-Total number of Concepts of domain ontology.

```
Procedure:
Do While i <= N
{
Remove HTML tags /*Special characters from Di.*/
Remove less meaningful terms or stop words.
Di = Di - S
Apply Stemming and find the root words.
w=w or prefix+w or w+suffix or Plural(w) or Tenseform(w)
$Di = (w_{i1}, w_{i2}, . . . , w_{im})$
  For j = 1 to m
  {
```

```
Calculate tf ($W_{ij}$) = Count($w_{ij}$, $D_i$)
tagf ($W_{ij}$) = Present($W_{ij}$, URL) + Count($W_{ij}$, Tags)
Get ontology entries Cjk containing Wij
Do while k <= C /* No. of concepts for $W_{ij}$ */
{
  tf ($C_{jk}$) = Count($C_{jk}$, $D_i$)
      tagf($C_{jk}$) = Present($C_{jk}$, URL) + Count($C_{jk}$, Tags)
Save $W_{ij}$ and its set $C_{jk}$ for $D_i$ along with their scores in the Database Table.
Repeat until i <= N
  {Repeat until j <= m
    {If $w_j$ presents in Di
      df ($w_j$) = df ($w_j$) + 1}}Do while j <= m
    {idf ($w_j$) = logN/df ($w_j$)
      Store df and idf for $w_j$ in the Database Table.}
```

The collection of web documents is preprocessed by removing stop words and performs stemming. This results in a set of pure words for each document. For each word the term frequency is calculated in the content, in the URL address and also in specific tags such as Head, Title, Link, Anchor, and etc. Then each word is mapped with the concepts of ontology. For each concept words the same concept frequency is calculated in the content, in the URL and also in the important tags. These values are indexed in the database table for the purpose of retrieval. In the last part of the algorithm the document frequency for each word and its inverse document frequency are calculated.

5. Discussion of common issues

We have discussed a preliminary of the existing and dynamic area in intelligent semantic search engines and methods.

a) Low precision and high recall
Some intelligent semantic search engine cannot show their significant performance in improving precision and lowering recall.

b) Identity intention of the user
User intention identification plays an important role in the intelligent semantic search engine. For example, introduced method for analyzing the requested terms to fit user intention and service provided will be more suitable for each user.

c) Inaccurate queries
We have user typically domain specific knowledge and user don't include all potential synonyms and variation in the query.

6. Conclusions

Semantic web is the future of Internet and semantic web is expected to re write the internet as we know it and change the way we search information on net. The searches will become personalized and the results will be more accurate and more relevant. The use of Semantic search on Resource Description Format with ontology relationships will help in the advent of this technology. Although there are many challenges that have to be overcome in order to do so but the possibility of this technology overcoming and replacing the traditional web model seem bright currently.

The traditional model of internet does not allow for intelligent searches and takes a lot of time because of the irrelevant searches being displayed too. Semantic Web can overcome all these problems to provide a better and rich user experience to consumers all over the globe. The next generation of web will better connect people and will further advent the information technology revolution.

Reference

[1] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, Scientific American, May 2001.

[2] Aleman-Meza, B., Arpinar, I. B., Nural, M. V. & Sheth, A. P. (2010). Ranking documents

semantically using ontological relationships. In Proc of IEEE fourth international conference on semantic computing (ICSC) (pp. 299–304).

[3] Lambert, F., Shanna, A. & Demartini , C. (2009). A relation based page rank algorithm for semantic web search engines. In IEEE Trans On Knowledge and Data Engg (vol 21(2), pp.123–136).

[4] Shah, U., Finn, T., Joshi, A., Cost, R. S. & Mayfield, J. (2002). Information retrieval on the semantic web. In: Proceedings of 10th international conference on information and, knowledge management [November].

[5] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., et al. (2004). Swoogle: A search and metadata engine for the semantic web. In: CIKM'04 (pp. 652–659).New York, NY, USA.[November].

[6] Stojanovic, N., Studer, R., Stojanovic, L. (2003). An approach for the ranking of query results in the semantic Web. In Proceedings of second international semantic web conference, (ISWC'03) (pp. 500–516).

[7] Ning, X., Jin, H., & Wu, H. (2008). RSS: A framework enabling ranked search on the semantic web. Information Processing and Management, 44(2008), 893–909.

[8] Wei, W.,Barnaghi, P., & Bargiela, A. (2011).Rational research model for ranking semantic entities. Information Sciences, 181(2011), 2823–2840.

[9] Koohler, Jacob, Stephen, Michael, & Rüuegg, Alexander (2006). Ontology based text indexing and querying for the semantic web. Knowledge Based Systems (19, pp. 744–754). Elsevier, Science Direct.

[10] Singh, Ramesh, Dhingra, Dhruv, & Arora, Aman (2010). Web search Engine using semantic taxonomy. IEEE Potentials, 36–40.