Semantic Structure Representation of HTML Document Suitable for Semantic Document Retrieval

Nidhi Tyagi
Shobhit University,
Meerut,India.

Rahul Rishi
Maharishi Dayanand
University, Rohtak, India.

R.P. Agarwal
Shobhit University,
Meerut, India.

ABSTRACT

The information on the WWW is available in various formats. The RDF and XML representation provides semantic knowledge about the document where as HTML mark-up only indicates the structure and lay-out of documents, but not the document semantics. The representation of the HTML document to semantic form can facilitate the extraction of knowledge from these documents in a more efficient manner. This paper proposes a technique for providing semantic structure to the HTML documents and stores it in the knowledge base as predicates, which helps in the retrieval of context related documents.

General Terms

Knowledge, information retrieval, semantic web.

Keywords

Semantic representation, contextual data, HTML, XML, knowledge base, predicate.

1. INTRODUCTION

The World Wide Web is the greatest repository of information ever assembled by man. Due to the sheer volume of available information, it is becoming increasingly difficult to locate useful information. The main obstacle is the fact that the Web was not designed to be processed by machines. Although, web pages include special information that tells a computer how to display a particular piece of text or where to go when a link is clicked, they do not provide any information that helps the machine to determine what the text means. Accessing and extracting semantic information from web documents is beneficial to both human and machines. Human can browse and retrieve documents in a semantically manner whereas machine can easily process such structured representations.

The Semantic Web is an extension of the Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [1]. From the architectural viewpoint, the Semantic Web is the evolution of the World Wide Web in such a manner as to maintain the structure and accessibility that exists currently, while adding new features that adhere to the architectural design principle of the Web effort. The W3C Semantic Web Activity Statement [2] also includes the following explanation of the Semantic Web: *The Semantic Web is a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications.* Tim Berner-Lee the director of W3C, described it simply as "a Web of data that can be processes directly or indirectly by machine [3].

The data is available in various formats like RDF/XML/HTML on the net. The RDF and XML representation, provides semantic knowledge about the document where as HTML mark-up only indicates the structure and lay-out of documents, but not the document semantics. As a result web documents are difficult to be semantically processed, retrieved and explored by computer applications. These documents can be processed and given an XML form, representing the concepts, relationships, URL of document selected.

A *semantic network* or *net* is a graphic form to represent knowledge in patterns of interconnected nodes and arcs and is an alternative to predicate logic as a form of knowledge representation. Thus knowledge can be stored graphically, with nodes representing objects in the world, and arcs representing relationships between those objects. Such representation is closer to the way humans' structure knowledge by building mental links between things than the predicate logic. Semantic–net can be represented in the form of predicate logic and stored in the knowledge base. On the bases of the keywords and the relationship existing between them, the documents are clustered contextually.

2. RELATED WORK

Lot of research is under progress in the field of semantic representation of documents. Some of the papers have been discussed in this section.

Terje Brasethvik and Jon Atle Gulla[4], proposed a technique for semantic retrieval of documents geared towards document management. For the construction of a domain based model referred to by the document repository, the researchers have used conceptual modeling approach and a semantic modeling language. This domain model is actively used for the tasks of document classification and search. Linguistic techniques are also the part of both the construction of the model and its use.

Comfort T. Akinribido, Babajide S. Afolabi , Bernard I. Akhigbe and Ifiok J. Udo [5], have presented a fuzzy-ontology based information retrieval system that determine the semantic equivalence between terms in a query and terms in a document by relating the synonyms of query terms with those of document terms. Hence, documents could be retrieved based on the meaning of query terms. The results presented show that the Fuzzy-Ontology Information Retrieval system successfully retrieve relevant documents to user's query, irrespective of different meaning and varieties of domain..

Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java and Yun Peng[6], has proposed SWOOGLE: Searching the knowledge on the semantic web. It uses the following components, the discovery components which automatically discover and revisited semantic web documents (SWDs) using a set of integrated web crawler, the digest component, which computes metadata for SWDs and semantic web terms (SWTs). The analysis component, it uses cache SWDs and the metadata to classify ontologies among SWDs and the service component that support both human and software agents through convention web interface and SOAP web server APIs.

Marut Buranarach[7], has suggested a architecture of deduction system for the semantic web consisting of three major components: Information Acquisition: gathers the Semantic Web data available in the RDF format. It transforms the data into a form that is suitable for the knowledge base. The second component is the knowledge base: allows automated deduction to be made over the acquired information and finally the knowledge retrieval: component allows the gathered information and new conclusions produced by the knowledge base to be retrieved and utilized. The critical look at the available literature reveals that there is a requirement for a technique that:

- The retrieved documents should be represented in the semantically, to depict the concept and relationships existing.
- Provides a more efficient way to satisfy users search query requirements.
- The semantic form of the inputted search query should be generated for faster and appropriate retrieval.

3. PROPOSED WORK

The data is available in various formats like RDF/XML/HTML on the net. The RDF and XML representation provides semantic knowledge about the document where as HTML mark-up only indicates the structure and lay-out of documents, but not the document semantics. As a result web documents are difficult to be semantically processed, retrieved and explored by computer applications. The proposed technique segregates the documents on the bases of their format and processes them separately. After segregation the XML format of the HTML document are formed which further help in the task of predicates generation of these semantic documents. Figure 1 below represents the architecture of the context based semantic information retrieval system.

3.1 Main components of the architecture

Crawl Agent: It downloads the .TOL and robots.txt file for resolved URL and segregates the internal and external links. It starts downloading pages for the URL in the queue. The RDF/XML page are forwarded for the predicate generator .and the hyper text URLs are submitted for converting them into XML format.

XML converter: The preprocessed HTML document, are transformed to the XML format of the document is generated with the tool Light HTML to XML converter [8].

Predicate Generator: the XML format of the document is treated for the natural language analysis and semantic analysis techniques to extract the semantic relationships between the selected concepts [9]. The semantic net form of the documents retrieved is created and this information is stored in the form of predicates in the knowledge base.

Matcher: The task of the matcher is to generate the predicate of the search query inputted by the user and find its corresponding match in the knowledgebase.

Knowledge store: Stores all the semantic representation of the documents crawled in the form of predicates.

Context identification: the association among the various keywords of the documents can be accessed by the predicates generated and the contextual identification can be done with the help of the on-line dictionary.

3.2 Proposed algorithm

The complete process is divided into three main tasks:
The *Information collection* component gathers the Semantic Web data available in the RDF/XML format and HTML format. This data is transformed into a form that is suitable for the knowledge base by the component *knowledge discovery* it also allows automated deduction to be made over the acquired information. The *information acquisition* component allows the crawler to gather URL's and download the pages and semantic information

The steps involved to complete the procedure are as follows:

Step1. Crawl agents download the documents for the set of seed URLs available, and store them in the repository.

Step2. The documents are segregated on the bases of their type (RDF/XML or HTML).

Step3. HTML documents are converted to XML form by convertor.

Step4. All documents in XML format are submitted to the semantic information extractor module, where the semantic net of the document is formed and stored in the knowledge store in form of prolog predicates.

Step5. The node value and the relationship depicted by the edges in the semantic net help us to identify the context of the document.

Step6. The indexer indexes the documents according to their semantic information and contextual sense.

Step7. Now when the user fires the query it is converted to predicate form.

Step8. The corresponding match for the generated predicate is found from the knowledge store.

Step9. Finally the URLs with the exact contextual match to the search query are made available.

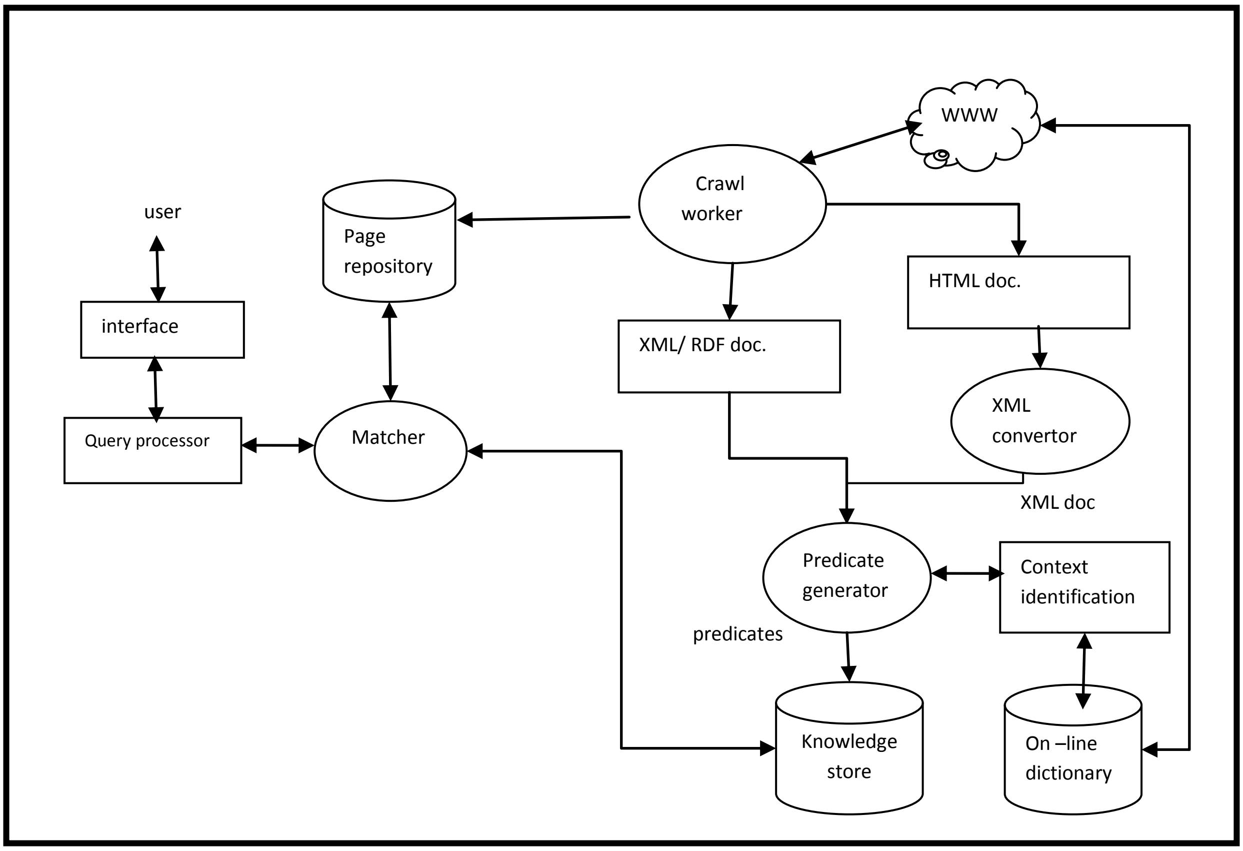


Figure 1: Architecture of the context based semantic information retrieval system

4. IMPLEMENTATION

The figure 2 represents the document crawled. The document is related to the keyword "**current**".

URL: http:// Electric_current.html.
TITLE: Electric current
KEYWORDS: electric, current, charge, conductor, ampere, ammeter, Columbus, ions.
DISCRIPTION:Electric current is a flow of electric charge through a medium. This charge is typically carried by moving electrons in a conductor such as wire. It can also be carried by ions in an electrolyte, or by both ions and electrons in a plasma.
The SI unit for measuring the rate of flow of electric charge is the ampere, which is charge flowing through some surface at the rate of one coulomb per second. Electric current is measured using an ammeter.

Figure 2: The document crawled for the word "current"

The XML format of the preprocessed HTML document with the URL address http://Electric_current.html, is represented in the figure 3. Similarly other documents related to the word 'current' but having different referential meaning are also crawled and transformed to XML format and are categorized into different clusters on the base of their contextual sense.

```
<title> electric current </title>
<keywords> electric, current, charge, conductor,
ampere, ammeter</keywords>
<Url> http:// Electric_current.html</Url>
<electric>
    <flow> charge</flow>
   <carried> electrons</carried>
  <measure> ampere</measure>
  <use> ammeter</use>
</electric>
     ------
      -------
```

Figure 3: XML form of document http://Electric_current.html

The semantic-net represented, integrates the documents on the base of synonyms, which can well establish the contextual sense of the various documents available on the WWW. Semantic –net can be represented in the form of predicate logic and stored in the knowledge base .On the bases of the keywords and the relationship existing between them, the documents are clustered contextually. The semantic net of the related documents is represented as shown in the figure 4.

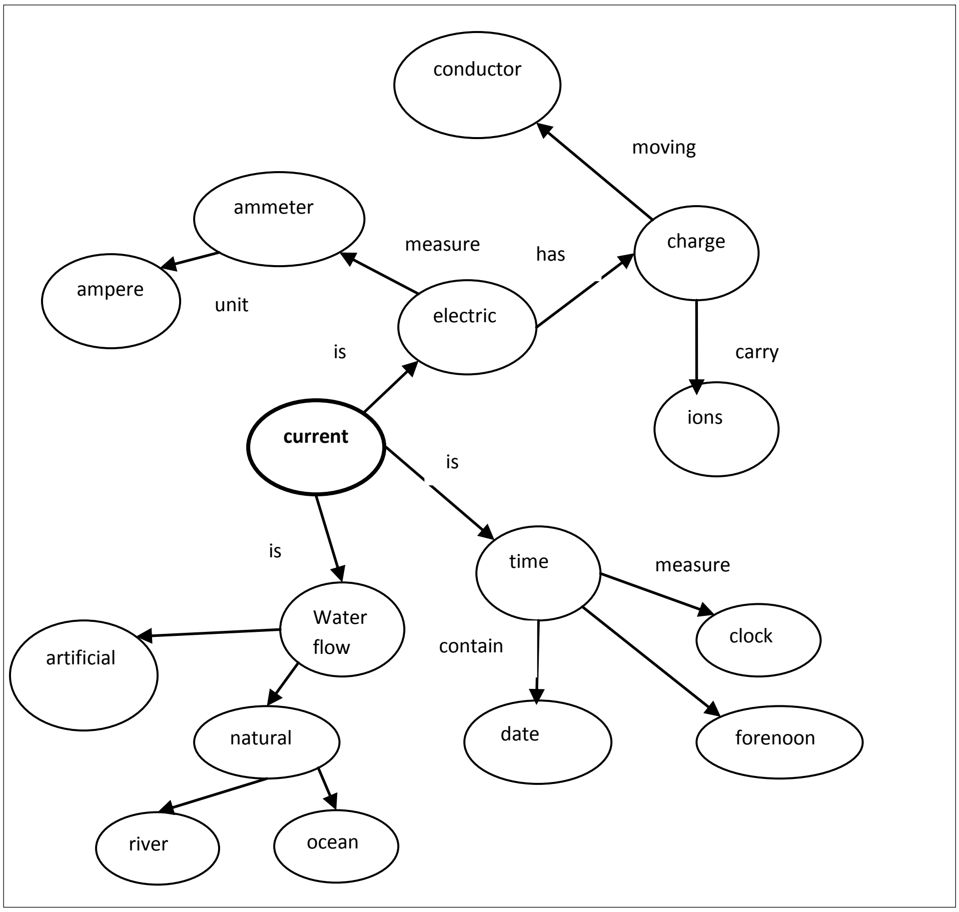


Figure 4: Semantic net of the documents related to "current"

The semantic net information is represented in the form of predicates, and the association of the predicated related to the particular document helps in extracting the contextual sense of the document along with their related URLs, as shown in the figure 5. The complete information of the knowledge-base is stored in form of prolog predicates. Figure 6 represents the predicates generated.

When the user fires the query it is transformed into predicate form. Corresponding match is identified in the knowledge base, and as the goal is achieved for the same as shown in the figure 6, the related URLs are made available to the searcher. The search interface for the context based semantic retrieval system is represented in figure7

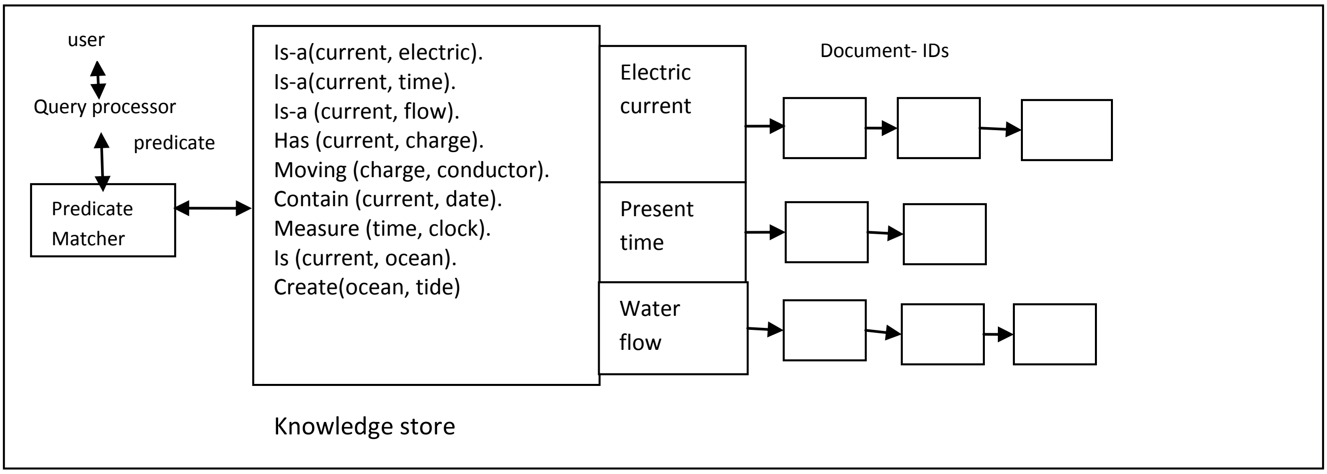
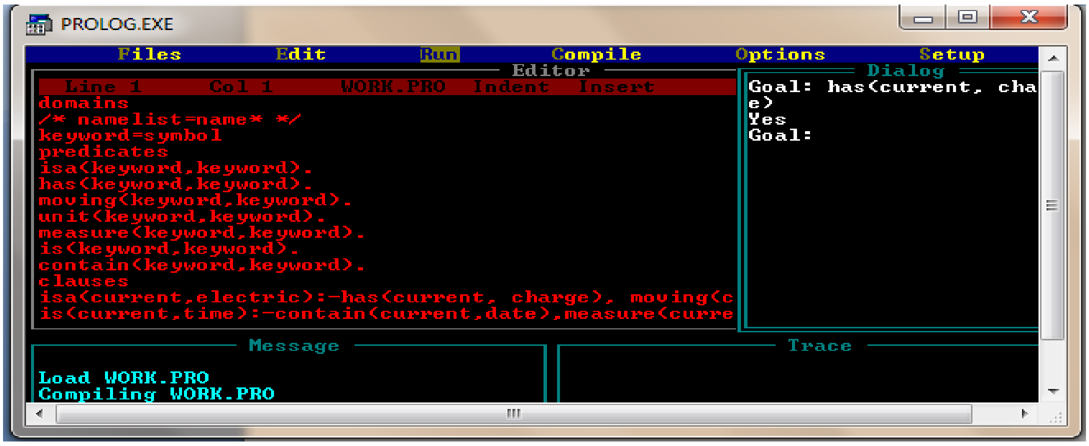


Figure 5: Representation of predicate matcher with knowledge store



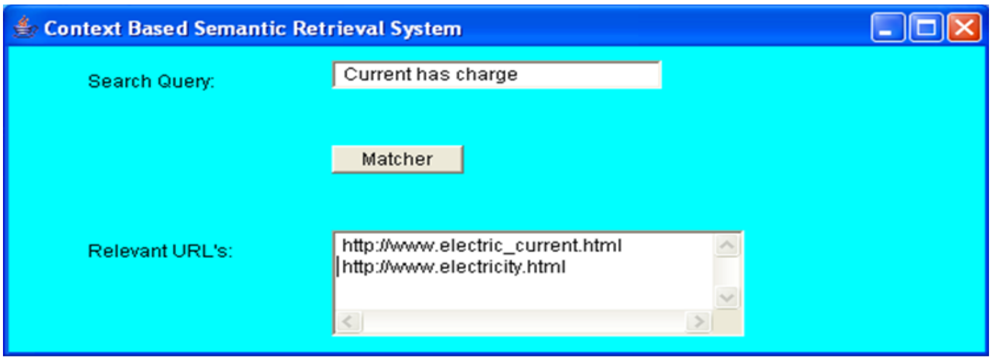
```
Line 1       Col 1       WORK.PRO   Indent   Insert
domains
/* namelist=name* */
keyword=symbol
predicates
isa(keyword,keyword).
has(keyword,keyword).
moving(keyword,keyword).
unit(keyword,keyword).
measure(keyword,keyword).
is(keyword,keyword).
contain(keyword,keyword).
clauses
isa(current,electric):-has(current, charge), moving(c
is(current,time):-contain(current,date),measure(curre
```

```
Goal: has(current, cha
e)
Yes
Goal:
```

```
Load WORK.PRO
Compiling WORK.PRO
```

Figure 6: Predicates representing the documents related to word "current"

.



Search Query: Current has charge

Relevant URL's:
http://www.electric_current.html
http://www.electricity.html

Figure 7: The search interface for context based semantic retrieval system

5. CONCLUSION

Semantic representation of the document plays an important role in supporting the task of document classification and identification. In this paper we have proposed a technique for providing semantic structure to the HTML documents and store it in the knowledge store as predicates. The proposed technique improves the performance of the searching system in terms of accuracy and efficiency for retrieving more, appropriate documents as per the user's requirements. As the documents are stored contextually the relevant URLs are made available to the user in short span of time

6. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lasilla, " The SemanticWeb", Scientific American, May 2001.

[2] W3C. Semantic Web activity statement. W3CTechnology & Society Domain Activity Statement,Available: http://www.w3.org/2001/sw/Activity. 2002.

[3] Berners-Lee, T., Weaving the Web: The original design and ultimate destiny of the WorldWide Web New York, NY: HarperCollins, 2000.

[4] Terje Brasethvik and Jon Atle Gulla ,"A ConceptualModeling Approach to Semantic Document Retrieval", Advanced Information Systems Engineering, 14th International Conference, pp.167-182, 2002.

[5] Comfort T. Akinribido, Babajide S. Afolabi , Bernard I Akhigbe and Ifiok J. Udo," A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback", International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[6] Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java and Yun Peng, "Swoogle: Searching for knowledge on the Semantic Web"University of Maryland Baltimore County, Baltimore.

[7] Marut Buranarach ," A framework for the organization and discovery of information resources in a www environment using association, classification and deduction", December 13,2004.

[8] Tool: Light HTML to XML converter.

[9] Sekine Proteus Project - Apple Pie Parser, http://nlp.cs.nyu.edu/app (Corpus based Parser) 2006.

[10] Parul Gupta and A.K.Sharma," Context based Indexing in Search Engines using Ontology", International Journal of Computer Applications,Volume 1 No. 14, pp 49-52, 2010.

[11] Nidhi Tyagi, Rahul Rishi and R.P.Agrawal," Context based Web Indexing for Storage of Relevant Web Pages", International Journal of Computer Applications (0975 – 8887) Volume 40– No.3, February 2012.

[12] Pell, "POWERSET - Natural Language and the Semantic Web". The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007.