XML-based Information Retrieval on the Semantic Web

Thet Zun; U Than Myo Naing
Computer University(Hinthata),Myanmar
thetzunlay10@gmail.com;

Abstract *The growth of the internet and technologies are useful for information search and retrieval on the Web. There are two methods of finding and interesting information such as; browsing which is suitable for situations where the goal is general information on the topic and querying or keyword based search which is appropriate when user has a clear goal. Although search engine technology has improved in recent years, there are staying many types of searches that return unsatisfactory results. This situation can be greatly improved if web pages use a semantic markup language to describe their content. In this system, keywords are first analyzed from the user to query find relevant pages are converted into Extensive Markup Language (XML) by using term database. Finally, it retrieves the ranked pages from entire domain area with the standard query results.*

Keywords: Keywords search, Semantic Web, Text Extraction

1. Introduction

Information retrieval technology has been central to the success of the Web. For the semantic web technologies to have an impact, they will have to be compatible with Web search engines and information retrieval technology in general. The Semantic Web has lived its infancy as a clearly delineated semantic web content will exist in separate documents that reference and describe the content of conventional web documents. With the vast expansion of the World Wide Web during the last few years the integration of heterogeneous information sources has become a hot topic. A solution to this integration problem allows for the design of applications that provide a uniform access to data obtainable from different sources available through the Web. Most computers now participate in the global Internet and the tasks we expect them to perform have changed accordingly.

2. Motivation

The Semantic Web has lived its infancy as a clearly delineated body of Web documents. When the desired Semantic Web document was not at hand, one was more likely to use a telephone to find it than a search engine. Current Web search techniques are not directly suited to indexing and retrieval of semantic markup. Most search engines use words or word variants as indexing terms. When a document written using some flavor of SGML is indexed, the markup is simply ignored by many search engines. Because the Semantic Web is expressed entirely as markup, it is thus invisible to them. Web search engines typically rely on simple term statistics to identify documents that are most relevant to a query. One vision of the Semantic Web is that it will be much like the Web we know today, except that documents will be enriched by annotations in machine understandable markup.

3. Related works

The web is currently a distributed mass of simple hypertext documents. WebKB [12] is a tool that interprets semantic statements stored in web-accessible documents. WebKB advocates the use of Conceptual Graphs and simpler notational variants for ontology and control commands that enhance knowledge readability and let its users combine lexical, structural, and knowledge-based techniques to exploit or generate web documents. Quest [3] was designed and implemented for querying and manipulating documents written in the OHTML [11] markup language. Quest uses the W3Lorel query language, based on the Lorel [1] language to query the OEM objects (semantic view), as well as the hypertext view (HTML tags) of the document. OHTML supports fine granularity semantic tagging of HTML pages. Quest uses the W3Lorel query language, based on the Lorel [1] language to query the OEM objects (semantic view), as well as the hypertext view (HTML tags) of the document. ELIXIR [4], an Expressive and Efficient Language for XML Information Retrieval, extends the query language XML-QL [7] with a textual similarity operator. XYZFind (Egnor & Lord) is a system for structured information retrieval using XML. It incorporates techniques for exploiting semantically structured XML to increase precision and recall and an extension to the classic inverted index to support structured Information Retrieval.

4. System overview

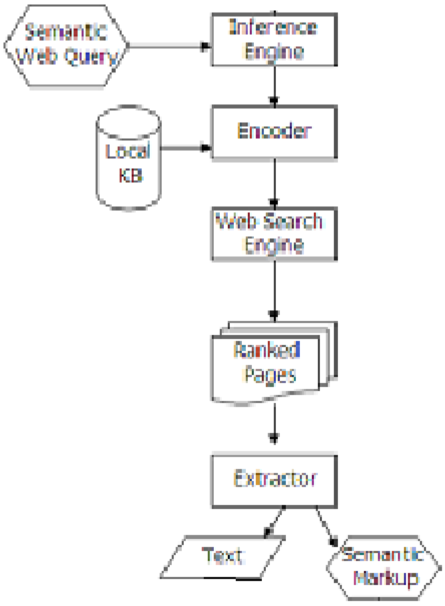

Figure 1: Design for Information Retrieval System

Step1: The system supports both retrieval-driven and inference-driven processing. Text is not useful during inference. To the extent that it is possible to automatically convert text to a semantic representation, such resulting representations can be used during inference. However, semantic interpretation is difficult at best, and unsolved in the general case.

Step 2: We must first encode the semantic markup query as a text query that will be recognized by a search engine. The query is submitted to one or more Web search engines. The encoder need to create consistent knowledge base. Data provenance is a term used for modeling and reasoning about the ultimate source of a given fact in a database or document.

Step3: Web search rely on today's broad coverage, text-based retrieval engines.

Step4: Search engine leaves us with a decision about whether to look into the ranked pages as a result set for more information to use in reforming.

Step5: The Extractor provides the text pages and semantic markup language. As a finial, the system result the retrieved information on the semantic web.

4.1. Semantic Web Integration

The World Wide Web Consortium (W3C) is promoting the Resource Description Framework (RDF). Whereas the web as we know it can be thought of as an ocean of pretty, linked, human-readable data islands, and XML vocabularies provide certain archipelagos with specific and incompatible dialects, the semantic web proposes to create a global sea of rich machine-comprehensible information. An integrated information base would allow uniform queries over information from all sources, and greatly expand the amount of knowledge that could be inferred from the combination of models. Semantic web research lies at the confluence of a number of other research streams. It borrows from work in databases, artificial intelligence, distributed systems, information theory and philosophy, among others.

4.2. Web search engines.

Web engines support users' information exploratory task particularity if there are not familiar with the domain. There is present the task action and information objects. It should be noted that the combination of task actions and information objects is not a one-to - one mapping. In this reason the relevant information is dependent on the task context. Each retrieved information provides links for description.

4.2.1. Information retrieval and Searching. Information retrieval (IR) is concerned with the process involved in the representation, storage, searching and finding of information which is relevant to a requirement for informatioin desire a human user. Querying is keyword based search and the search engines post the user query to their index of keywords and return a ranked list of documents. IR on the sematic web can be addressed according to three different points of view: there are developers of ontologies focusing on the representation of domain knowledge, annotation of web resources creating semantic annotation and queries for searching the web. The goal of the semantic web is to identify more web-based data and their interrelationships so that searches can be more effective.

5. Converting HTML to XML

Most of the information on the Web today is in the formed of Hypertext Markup Language (HTML) documents which are viewed by human with a browser. HTML documents is designed for presentation purposes, not automates retrieval, and the fact that most of the HTML contents on the Web is ill-formed ("broken").In the future some if not most Web contents may be available in formats more suitable for automated, in particular the Extensible Markup Language (XML). XML has become absolutely essentials for enabling data interchange between other wise in compatible systems. The volume of XML contents available on the Web today is still miniscule compared to that of HTML. It is therefore reasonable to study ways of translating existing HTML contents to XML.

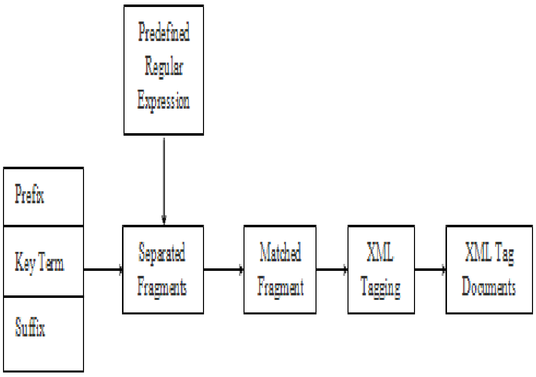

Figure 2: Process of HTML to XML Conversion

5.1. Extensible Markup Language (XML)

The Extensible Markup Language (XML) is a set of rules for defining textual formats for structured data storage. It comes from a tradition of markup languages (such as HTML); themselves designed according to the rules of XML's precursor SGML. XML's rules are a vast simplification of SGML, which is widely believed to be responsible for XML's widespread popularity. In this thesis, I take the less ambitious route of only integrating the basic structure of XML documents without trying to decode its meaning. An application with dialect-specific knowledge can then transform this structural model into the information actually contained in the document.

5.1.1. Basic Structure of XML Documents. All XML documents follow a basic tree structure of nested elements intermixed with free-form text. Every element also has any number of attributes, each associating a simple string value to the element. Both elements and attributes are labeled. Based on this description, XML documents are normally represented as a labeled tree.XML elements are nests with list characteristics (totally ordered and accepting duplicates). Elements contain other elements and strings. Attributes are binary relationships between the element and the assigned string value.

Example of XML documents
<person id ='1'>
<firstName>Emily </firstName>
<familyName>Lu</familyName>
<date>15-1-2001</date> </person>

6. System architecture

The semantic web is to enhance the ability of both people and software agents to find documents, information and answers to queries on the Web. To explore the tight integration of search and inference, a framework designed to meet the following desiderata:

- The framework must support both retrieval-driven and inference-driven processing.
- Retrieval must be able to use words, semantic markup, or both as indexing terms.
- Web search must rely on today's broad coverage, text-based retrieval engines.
- Inference and retrieval should be tightly coupled; improvements in retrieval should lead to improvements in inference, while improvements in inference should lead to improvements in retrieval.

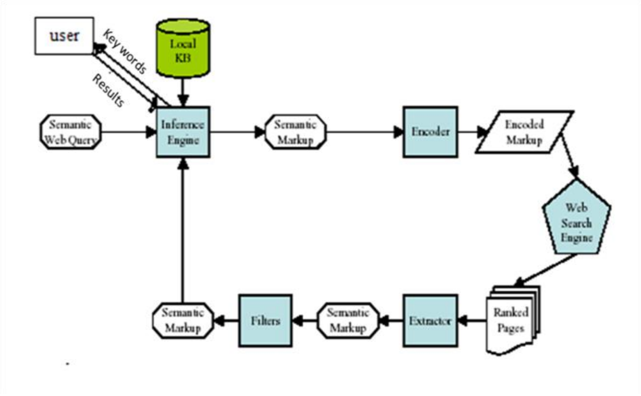

Figure 3: System Architecture

7. Experimental results

If the user's goal is retrieval, this might simply be semantic markup encoding the concepts being sought (*e.g.,* using XML-QL [10] or XIRQL .Alternatively, if the goal is inference, the query might be a statement the system is to prove. In either case, the query is submitted to the inference engine. For retrieval, the inference engine may choose to perform limited forward chaining on the input (as a text retrieval engine might perform thesaurus expansion). For proof, the inference engine will generate a partial proof tree (or more accurately, one in a sequence of partial proof trees), using its local knowledge base to the extent possible. The inference engine produces a description of the semantic markup to be sought on the Web.

Metadata of documents is useful for many kinds of documents processing such as searching, browsing and filtering. The metadata are even more important than the contents because users always query or browse information by metadata (for example, searching by title keywords).

We have been able to improve the precision of our search through the use of controlled vocabularies or limit the searches to the descriptor, identifier, author, title, or source fields for many years. The metadata, if well chosen, should describe the central topics of a documents. Thus, it should be given a high weight, relative to the appearance of those terms in the full test of the documents.

Some sort of inference engine, to identify facts and rules needed by the inference engine to reach its desired conclusions, to search the Semantic Web for such facts and rules, and to incorporate the results of the search into the inference process. If the user's goal is retrieval, this might simply be semantic markup encoding the concepts being sought (*e.g.*, using XML-QL [10] or XIRQL [15]). In either case, the query is submitted to the inference engine. We want to use a traditional Web search engine for the retrieval, we cannot simply use the output of the inference engine as a search query. Rather, we must first encode the semantic markup query as a text query that will be recognized by a search engine. The query is submitted to one or more Web search engines. The result will be a ranked list of Web pages, which either contain semantic markup themselves, or refer to companion pages that do. Some number of these pages must be scraped to retrieve their semantic markup. Control over how many pages to scrape, and over whether to scrape additional pages or to issue a new Web query, resides with the inference engine.

8. Conclusion

The Semantic Web will contain two kinds of documents. Some will be conventional text documents enriched by annotations that provide metadata as well as machine interpretable statements capturing some of the meaning of the documents' content. Information retrieval over collections of these documents offers new challenges and new opportunities. We have presented a framework for integrating search and inference in this setting that supports both retrieval-driven and inference-driven processing, uses both text and markup as indexing terms, exploits today's text-based Web search engines, and tightly binds retrieval to inference. While using on-demand retrieval introduces a possible bottleneck when the facts needed by the inference engine must be retrieved together with the requested data. Integrating heterogeneous information sources is needed in order to provide a uniform access to data gathered from different sources available through the Web. The proposed integration architecture combines semantic metadata with on-demand retrieval.

9. References

[1] Abiteboul, S.; Quass, D.; McHugh, J.; Widom, J.; and Wiener, J. 1997. The lorel query language for semistructured data. International Journal on Digital Libraries 1 68–88.

[2] Arocena, G. and Mendelzon, A. 'WebOQL: Restructuring documents, databases and webs.' In International Conference on Data Engineering, pages 24-33. IEEE Computer Society, 1998.

[3] Bar-Yossef, Z.; Kanza, Y.; Kogan, Y.; Nutt, W.; and Sagiv, Y. 1999. Quest: Querying semantically tagged documents on the world wide web. In Proc. of the 4th Workshop on Next Generation Information Technologies and Systems, volume NGITS'99.

[4] Chinenyanga, T., and Kushmerick, N. 2001. Elixir: An expressive and efficient language for xml information retrieval.
In SIGIR Workshop on XML and Information Retrieval.

[5] Cost, R. S., Finin, T., Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., Chen, H., Kagal, L., Perich, F., Zou, Y., and Tolia, S. 'ITTALKS: A Case Study in the Semantic Web and DAML+OIL.' IEEE Intelligent Systems17 (1):40-47, 2002.

[6]DAML+OILDesign Rationale 2001.www.cs.man.ac.uk horrocksSlides/index.html.

[7] Deutsch, A.; Fernandez, M.; Florescu, D.; Levy, A.; and Suciu, D. 1999. Xml-ql: A query language for xml. In Proc. 8th Int. World Wide Web Conference.

[8] Egnor, D., and Lord, R. Structured information retrieval using xml. XYZFind Corporation, Washington, USA.

[9] Forgy, C. 1982 Rete: A Fast Algorithm for the Many Pattern/ Many Object Pattern Match Problem. Artificial Intelligence 19 17–37

[10] Fuhr, N., and Grojohann, K. 2000. Xirql: An extension of xql for information retireval. In SIGIR Workshop on XML and Information Retrieval.

[11] Kogan, Y.; Michaeli, D.; Sagiv, Y.; and O.Shmueli. 1997. Utilizing the multiple facets of www contents. In
3rdWorkshop on Next Generation Information Technologies and Systems, volume NGITS'97.

[12] Martin, P., and Eklund, P. 1999. Embedding knowledge in web documents. In Proceedings of World Wide Web Conference
(WWW8).