Information Retrieval on the Web: An Approach Using a Base of Concepts and XML

Abstract: The emergence of XML as a new standard for documents semi-structured on the Web opens opportunities to improve the process of interrogation of Web sites. HTML's major inconvenience is that it does not allow one to distinguish the logical and physical aspects of documents. We propose a model of data-web based on XML, and a concepts base composed of a meta-data base and a domain thesaurus. The meta-data base includes information on the content, the semantic structure, and the organisation of data of the site. The process of searching for information is based on the exploitation of the elements of the concepts base and allows for an interactive search for relevant documents (or extractions from documents)

1. Introduction

The World Wide Web (WWW) contains huge amounts of information that is available at web sites, but it is difficult and complex to retrieve pertinent information. Indeed, a large part of this information is stored as static HTML (HyperText Markup Language) pages that are only viewed through a web browser. A major inconvenience of HTML is that it allows, with difficulty, distinction between logical and physical aspects of the documents.

Different search engines (Alta Vista, Hot Bot, Lycos, Yahoo!, etc.) provide keyword-based search facilities to help users to retrieve information and the results again come as HTML pages. From one or several words, such engines collect a whole range of information from web documents that use these words, but results are often limited. The search engines do not take account of the semantics of the documents. This approach can be interesting for research on the whole web but becomes obsolete when used on a simple web site. The quality of the results can be improved with integration of meta-information in web sites, that is information on the content, the structure and the organization of the data of the site.

A first approach to meta-information integration consists of annotating the pages of an existing site with tags of meta-content. One creates a formal description of the content with HTML documents. Dobson and Burill (Dobson, 1995) are the first to propose a semantic marking of the document content; they relied on the Entity/Relationship model that allows one to capture the conceptual level of the site. They introduced a set of three tags that allows one to define the entities or concepts in documents while labeling the sections of the text as attributes of these concepts, and while defining relationships of an entity to another external entity. Ontobroker (Fensel, 1998) is an HTML extension that allows one to declare ontological annotations in the web pages. It uses some attributes-values couples in hyperlink tags. The authors of documents use these tags to delimit an element. WebKB (Martin, 1999) integrates some knowledge in the web pages while using knowledge representation languages to represent and to index information in the documents with HTML tags. It exploits the conceptual graph formalism and the ontology of concepts and kinds of relations.

In the second approach, the researchers proposed semi-structured data models, databases and languages to model, to store and to interrogate the web sites. The interrogation languages like W3SQL, WebSQL (Mendelzon, 1997) are interesting in the setting of precise information retrieval based on strings and structures (tags, link names...) retrieved in the documents. WebSQL introduces the navigable link notion to answer a query.
In other systems the query is done on an uniform structure constructed from the conversion/integration of heterogeneous data sources (Cluet, 1997), notably the web, while using a semi-structured data model. The OEM model (Object Exchange Model) (Abiteboul, 1997) which is detailed in (Gardarin, 1999) has been conceived for semi-structured data management. In OEM the data constitute a graph with an object unique root. The YAT system (Cluet, 1997), (Simeon, 1998) is based on a data model and a declarative language for integration and conversion of heterogeneous data sources. This model is more powerful than OEM: on the one hand the data are represented in YAT by arbitrary graphs, and, on the other hand, part of it has the capacity to possibly represent some data with their type.

In all these systems, a web site is seen as a set of HTML pages. Our approach is based on the use of new emergent standards around XML (eXtensible Markup Language). In the process of information retrieval, we exploit two techniques which are very familiar to the users of the Web: i) initialization of the process of search by simple requests by means of key words; and, ii) exploitation and improvement of obtained results.

The information representation that is based on the document’s logical structural integration with its semantic content allows the exploitation of the first technique. It uses a concepts base of the domain that includes a thesaurus and meta-data information contained in the data-web.

In classic information retrieval systems, keywords enable finding a list of documents containing these words. In our approach, the pertinent information (documents, simple parts of documents simply or automatically composed) is found from structural elements of the document (the tags), from the text described by these elements, and also meta-information describing its semantic content.

The second technique used is from the results of an interrogation query. These results are presented on a hypertext card (interactive card of concepts) allowing the user to refine the results by navigation or interrogation in the space of solutions proposed.

After this introduction, section 2 is dedicated to the presentation of some main features of XML to allow the reader to understand the concepts explained in the following sections.

The information representation in the data-web is described in section 3. We present the technique that is used to integrate the semantic contents of the document with its logical structure and the concepts base composed of a meta-data base and a domain thesaurus.

We present in section 4, the interrogation process which is based on the combination of the logical and semantic structures of the documents and the exploitation of the DOM. Examples are extracted from the SIMEV application (Information System on the Senegal stream Valley's enhancements). This application is part of the GIRARDEV program that integrates into its actions the conception of a regional observatory of development. An implementation of this approach is currently under way in Java on a Java-enabled platform.

2. Data-web and XML
2.1. Introduction to XML

The aim of this section is to give some basic notion about the main features of XML in order to keep the reader able to understand the concepts that we introduce in the next sections. For more details, see Michard (1998).

XML has recently emerged as a new standard for representation and exchange of data on the Web (Bray, 1998). It is a subset of the standard SGML (Yoshikawa, 1996; Michard, 1998). This meta markup language defines its own system of tags representing the structure of a document explicitly. HTML presents information and XML describes information. Let's consider the piece of XML document below:
A document is composed mainly of a tree of elements that forms the content. XML documents are composed by markup and content. In the above example, the coupling of tags <HYDROLOGY> and </HYDROLOGY> is the markup for the string, important streaming.

Attributes are another important kind of markup; they are names-values couples that appear after the name of an element in the opening tag. For example, <SOIL CODE="DGX"> is element SOIL with an attribute CODE that has DGX value. A well-formed XML document doesn't impose any restrictions on the tags or attribute names. But a document can be accompanied by a DTD (Document Type Definition) which is essentially a grammar for restricting the tags and structure of a document. An XML document satisfying a DTD is considered a valid document.

A Document Object Model (DOM) defines the logical structure of the document (HTML and XML) and the way a document is accessed and manipulated by application programs. With the DOM one can build documents, navigate their structure, and add, modify, or delete elements and content. It is designed to be used with any programming language.

2.2. Data-web: an XML Architecture

A dataweb is a collection of data: (i) structured such as the ones stored in relational or object databases; and, (ii) semi-structured as the data of Web. A dataweb system is a system that manages a data-web.

The data-web model that we propose integrates the domain knowledge represented into a concepts base (Brou, 1997; Hocine, 1999). Its general architecture is defined in Figure 1.

![Data-web Architecture](image)

Figure 1: Data-web architecture

The process of information retrieval associated with this model exploits two techniques which are familiar to Web users: i) to begin searching with a simple query, usually with keywords; and, (ii) to exploit and to refine the results.
The classic information retrieval systems use a thesaurus whose main function specifies the semantic content of used terms during the stages of indexing and searching. The concepts base integrates a thesaurus of the domain and meta-data base describing the semantic contents of the data-web. In fact, a data-web is defined as a set of XML documents which represent the web site data and/or the results of queries on databases.

Figure 2 presents the conceptual Entity/Relationship schema of the data-web. The association "is included" represents the inclusion of an element in another. Information contained in an XML document is accessible through the DOM by exploiting tags and the contents of elements. These elements are therefore potential answers to a query or interrogation. To this logical structure we associate a semantic structure (expressing semantic contents) which acts to:

(i) characterize the descriptive information or external attributes;
(ii) define semantic elementary or compound units;
(iii) index the content of the document by keywords come from the thesaurus.

Figure 2: Conceptual schema of a data-web

3. A Concepts Base to Improve the Research Process in a Data-web

The thesaurus and the meta-base are the main constituents of a concepts base.

The thesaurus: A thesaurus is a set of concepts connected by hierarchical relations, of equivalence of association (Boughanem, 1992). These semantic relations are represented in the conceptual model, above, by association "conceptual relation". The hierarchical relations define the notions of generalization and specialization. They find a document which treats very specific concepts if one asks a question at a more general level. Inversely, to a very specific question, this type of relationship finds, if need be, more general documents. For example, concept "Enhancement technique" is a sub-concept of "enhancement". The notion of equivalence (or synonymy) translates semantic equality between two concepts.

The notion of equivalence (or of synonymy) translates semantic equality between two concepts. For example, the concepts "arable zones" and "cultivable zones" are equivalent. The associative relations (or neighborhood) are used to cover strictly connected concepts (for example "enhancement" and "kind of soils"). They allow to clarify or to extend the sense of a question. The thesaurus is described in the XML format.
The meta-data base: Meta-data constitute a description of the XML documents. These last ones are represented by a meta-information set which is constituted of:
- descriptive information (title, date, author, etc.) or external attributes;
- a set of keywords (tag <Keyword>) making reference to the concepts of the thesaurus. It is the result of the indexation process of documents elements of XML base. We call this part of meta-data, semantic meta-data. This aspect is exploited during the interrogation process of the data-web. We can associate to every element of the document a semantic unit with the attribute SU. These semantic units regroup elements which have a meaning in the context where they are presented together.

![Diagram of concept definition and its XML representation in the SIMEV thesaurus](image)

Figure 3: Example of concept definition and its XML representation in the SIMEV thesaurus

![XML representation of meta-data](image)

Figure 4: XML representation of meta-data

4. A Research Process Based on Logical and Semantic Structure
The data-web model thus proposed allows interactive information retrieval. This research is done according to two approaches: interrogation, and navigation.

The research process is initialized by queries of simple interrogation by means of keywords which belong to the thesaurus (Figure 3). The semantic structuring of a document retrieves a semantically structured document, i.e., the same document but structured into semantic units instead of elements (logical units). It is on this document that the indexing is done; every semantic unit is indexed by a set of keywords (Figure 4).
For every interrogation, the information retrieval system compares the query with information stored in the dataweb. An interrogation query is treated in two stages: (i) normalization of the query; and, (ii) execution of the normalized query that provides a list of pertinent documents (or extracts of documents). A generalized query is a normalized query which is extended to the semantic relations of the thesaurus. For example, we complete the query by using the synonyms of the present keywords in the query. These keywords are directly accessible in the thesaurus (Figure 5).

The query normalization transforms, by means of the thesaurus and of the DOM, to obtain a query semantically homogeneous with the concepts base and syntactically correct with the DOM.

The results of this query are presented in a hypertext card. This interrogation interface (Figure 5) permits access at any time to the thesaurus which can be consulted.

The generalization of the query allows one to retrieve semantically interesting documents. The reformulation of query is not automatic but directed by users with invocation of knowledge contained in the thesaurus. The documents are re-stored according to their degree of semantic resemblance with the query.

In our example, if the query is "kinds of soils", all documents indexed by the concept "kind of soils" are provided by the method of questioning. But the generalization of this query, with the help of the thesaurus, takes out, again, the semantic links of this concept with "soils" (generic concept) and "enhancement" (neighboring concept). The documents indexed by concepts "soils" and "enhancement" are then provided in the answer.

![Figure 5: Interface of interrogation and navigation](image)

5. Conclusion

In this article, we showed the specificity of HTML applications and their limits in the search for relevant information. We proposed a conceptual model of a data-web based (i) on the document's logical structure integration with its semantic content, and (ii) on a representation of data in the XML language.

This approach to integrating meta-information improves the information research process, which is founded on the use of the DOM and Java language. Currently, we are working on another statistical method to improve the degree of semantic resemblance. The
specificity of environmental information systems (Dzeakou, 1998; Gayte, 1997) makes environmental observatories potential applications for our data-web model.

Notes
1. Interdisciplinary Group of research for support regional planning and local development: set by the regional Council of Saint-Louis (Senegal) with Senegalese and foreigner researchers and experts.

References


