Virtual WWW Documents: a Concept to Explicit the Structure of WWW Sites

Bich-Liên Doan (*doan@emse.fr)*
RIM Dpt., Ecole Nationale Superieure des Mines de Saint-Etienne
Saint-Etienne, France

Michel Beigbeder (*mbeig@emse.fr)*
RIM Dpt., Ecole Nationale Superieure des Mines de Saint-Etienne
Saint-Etienne, France

Abstract

This paper shows a new concept of a virtual WWW document (VWD), as a set of WWW pages representing a logical information space, generally dealing with one particular domain. The VWD is described using metadata in the XML syntax and will be accessed through a metadata.class file, stored at the root level of WWW sites. We'll suggest how the VWD can improve information retrieval on the WWW and reduce the network load generated by the robots. We describe a prototype implemented in JAVA, within an application in the environmental domain. The exchanges of such metadata lay in a flexible architecture based on two kinds of robots : generalists and specialists that collect and organize this metadata, in order to localize the resources on the WWW. They will contribute to the overall auto-organizing information process by exchanging their indices, therefore forwarding their knowledge each other.

Keywords: *information retrieval, WWW, search tools, metadata, cooperative architecture.*

1 Introduction

Information retrieval on the WWW has become one major research problem, partly because resources available on the WWW are heterogeneous, scattered and volatile. Another reason is the scalability problem due to the continuous growth of information volume. The WWW is viewed as a hypertext graph where the nodes are HTML pages that can be accessed through their URLs, and the links are textual anchors inside a page that point towards another URL and that enable the user to jump from page to page. HTML structures the display of pages, but provides very little information about the content and organization of pages or collections of pages; it is precisely this structured content we want.

To help users to find information on the WWW, search tools have been designed. Most of them use the inverted file technique to build indices and the boolean model to match the request with the indices of the database. The pages are indexed asindependent documents without any structure.

Besides the problem of noise and silence, the ranked list of hundred thousands of URLs given as responses is very difficult to exploit, because unless we download each HTML page, we have no idea of "is this page really tied with the subject of our question?".

In short, because the million pieces of information that build the WWW hypertext network are organized in a "flat" way, current WWW search tools are not capable of extracting the logical structure and the context around HTML pages.

To help to capture the structure and the semantics associated with each site, we propose to define the concept of virtual WWW documents as abstractions of sets of pages at different levels of granularity, which can be organized into different hierarchies of clusters in the sense of [11]. We introduce metadata that describe the VWDs precisely.

With regards to the limits of the WWW metadata, see [10], we experienced the use of metadata within a community of users, in a particular domain, the ”sustainable development”.

We propose to define a flexible architecture using the existing and available search tools on the WWW which will enable everybody to participate in the improvement of document descriptions. This architecture is based upon high interactivity between search tools and a progressive organization of information on the WWW. With minimal effort we can use cooperation between existing isolated elements of the WWW, resolving the problems of servers and bandwidth overload, scalability and context around information spaces.

In section 2 we give a model of what we call a Virtual WWW Document (VWD), we explain how to create and describe VWDs with metadata in section 3; then we describe the architecture of generalists and specialists which can take benefit from the use of metadata. Finally, we review related work and offer conclusions.

2 The Virtual WWW Document (VWD) model

We start with the following remark :
In the “printer” world, it is easy to capture the outline of a book, by reading the table of contents of the book, or its summary. On the WWW, the logical structure of pages is not explicit, even if the author of pages has a structured and hierarchical view of his information. To find out information without any ambiguity, we need to use a search word in a particular context, (for example we want to know that this set of pages is linked to environment, with environment meaning ecology and not computer environment).

In order to extract the context and organization of a set of HTML pages, we propose to define virtual WWW documents, we give a representation of them and we use metadata to describe them.

The metadata are composed of both pieces of information about the content of the document and external to that content. For instance, at the book level the author name, the publication date and the title are outside its semantical content.

Our system is based on two simultaneous level access : the data level (content of pages) and the metadata level (set of attribute-value pairs describing the context of pages).

2.1 Definitions

In this section we define the concepts and relations manipulated by the system : resource, VWD, relation, context, ontology, scheme, site.

Definition 1 *A HTML page is a WWW resource which is identified and accessed through its URL. The relation Page(URL, title, date, text) contains the title, last modification date, text.*

Definition 2 *A hypertext link is represented by a relation Link(source, label, target) where source is the URL of the resource the link start from, label is the text bound to the link and target is the URL of the destination page.*

We give now a definition of metadata :

Definition 3 *Litterally “data about data”, metadata is machine understandable information about WWW resources or other things. Metadata has a semantics and a structure about people, things, concepts or ideas.*

Definition 4 *An ontology is a domain specific metadata : ”objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them”.*
We define Ontology(name, ref), where name is a String identifying the ontology, for example ”thesaurus GEMET” in the environmental domain, ref is the URL of the service or database providing the vocabulary and relations between terms.

Definition 5 *A scheme is a data model used to represent the data. To describe the information in our system, we specify the set of elements of metadata, the syntax and the semantics of each element. Scheme(name, metadata) is identified by its name, a set of elements of metadata, for example the Meta Data Dublin Core.*

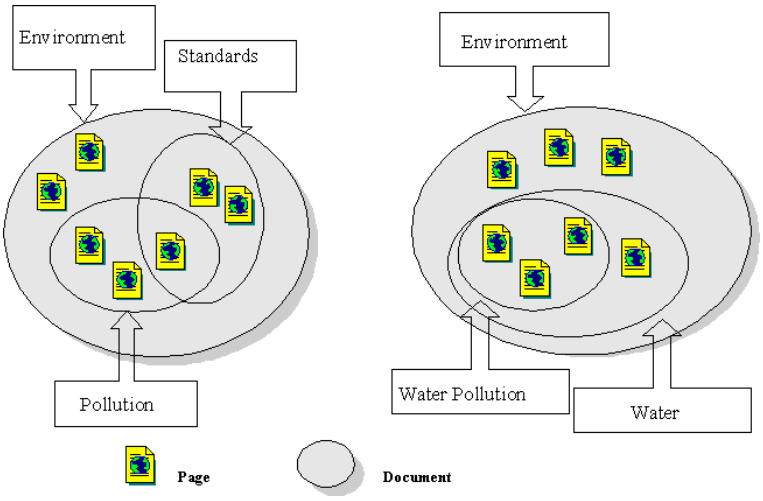

Figure 1: Several organizations of the same set of environmental pages

This scheme may be stored within the site, at the root place or may be a reference to a URL stored somewhere else.

Definition 6 *A relation is represented by Relation(name, type) where name is the name of the N:M relationship between VWDs, or VWD and HTML pages. Type is the type of the relationship.*

Examples of relations are association, child, contains, reference .... We use two important relations :
- child is the composition relationship with other subdocuments;
- contains is the set of HTML pages belonging relationship.

We introduce now the notion of context around a page or a document : Each page or document belongs to a context. Let's go back to the library example, and consider a page of a book. This page belongs to a section, this section belongs to a chapter which belongs to the book itself. Each level of organization of the book gives a context, from the highest level (the root is the book) to the next level in the hierarchy up to the page level. These contexts may inherit from their father context, depending if the attributes are dynamic descending (can be propagated down to the hierarchy). So we suggest to make explicit these contexts within the WWW documents, by giving some information about pages or set of pages contained within a document. The Fig. 1 shows how the WWW pages can be organized into clusters called VWDs, one page can belong to different documents, therefore bounded to different contexts.

Definition 7 *A context is represented by a tuple Context(name, Ontology, Scheme, $\{Field\}$), where name is a string uniquely identifying the context, a set of 4-tuple Field(attribute, qualifier, value, type) where :*

- *attribute is the name of the field;*
- *qualifier brings in more precision to the field;*
- *value is a set of keywords belonging to an Ontology or not;*
- *the type of value $\in \{String, Int, Float, Boolean, Date\}$.*

Definition 8 *Let us give a definition of a VWD :*
a collection of pages clustered according to some criteria, for instance pages dealing with the same subject like environment, representing a semantic information unit which can be given as a response to a user's request.
A VWD is represented by VWD(URL, Context, [Relation]) where :

- *URL is the identifier of the page or service describing the VWD;*
- *Context is the metadata linked to the VWD;*

- Relation is a typed link between VWDs, for example the link child enables to define a Directed Acyclic Graph (DAG) of VWDs.

Each page can belong to one or several documents. A document itself can be a component of other documents.

Definition 9 *A WWW site is characterized by Site(domain, {resource },[{VWD }], [metadata.class]) where :*

- domain is the domain name of the server hosting the WWW resources (for example www.w3c.org);
- resource is any data available on a WWW site through a URL;
- VWD is a collection of WWW pages or resources which can be explicited and organized into a DAG of documents;
- metadata.class file is the context of the VWDs available on a WWW site, and stored at the root level of the site, accessible through a URL.

The last two components of this relation are optional, we see later how they can be generated and maintained.

2.2 Describing a VWD with metadata

This semantics and structure may be different, according to the creator of metadata, and it would be naive to consider that only one standardized metadata will be used by every one. The recommendations given by PICS[3], RDF[4], including a meta-metadata to define the type of descriptions used will be adopted, in conjunction with the semantics chosen by OCLC/NCSA Workshop[16].

In the following, we give an example of our metadata. Our model requires the exchange of indices between different entities, in an extensible and flexible way. XML is a document description language [1], which enables the representation of structured document, and which is flexible and simple for the authors to use. Moreover, it will be translated by the next versions of "browsers" from Netscape and Microsoft.

So we opt for a XML syntax and we include the semantics of Metadata Dublin Core.

```
<?xml version="1.0"?>
<?xml:namespace ns="http://groseille.emse.fr/DC" Prefix="DC"?>
<!DOCTYPE DOCUMENT "http://www.emse.fr/~brodhag/projelev#d0.">
<DOCUMENT>
    <DC:Language>" fr-French" </DC.Language>
    <DC:Title> "environment" </DC:Title>
    <DC:Subject>
        <DC:scheme>
                <DDC> " 333.3" </DDC>
        </DCscheme>
    </DC:Subject>
    <DC:Subject> "pollution" </DC:Subject>
    <DC:Relation>
         <type="child">
                "http://www.emse.fr/~brodhag/projelev/RESOURCES"
         </type>
    </DC:Relation>
</DOCUMENT>
```

3 Creating, indexing and querying the VWDs

In this section, we introduce the problem of creating the VWDs, how these documents are maintained and by whom, finally we explain how to retrieve them.

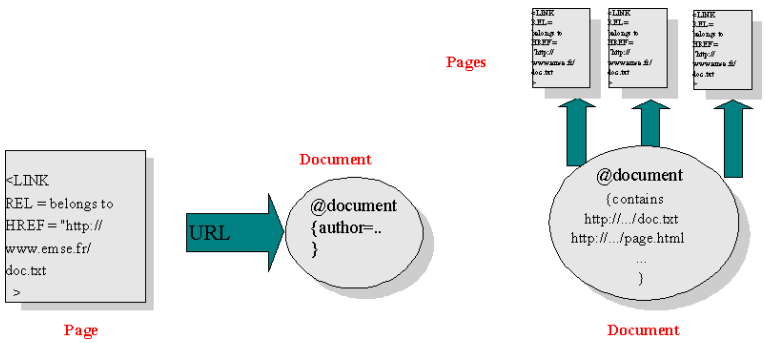

Figure 2: Creation of VWDs

3.1 Creation of VWDs and their context

Two solutions can be adopted to generate the VWDs. First the author of HTML pages add META tags inside its pages to indicate that these pages belong to one (or several) VWD located and described in a metadata.class file in the host server. For example he uses

```
<META NAME="containedIn"
CONTENT="http://groseille.emse.fr/metadata.class#d0>
```

This solution is practical for the administration of VWDs because if a page is deleted or modified, no change has to be done in the VWDs. If author are encouraged to put META keywords, description, subject inside the pages, then it will help to apply clustering algorithm to automatically build the hierarchies of VWDs.

A second consists in creating only the VWDs and using the relation contains to include the set of pages inside the VWD.

```
<DOCUMENT>
  <DC:Relation>
         <type="contains">
         "http://www.emse.fr/~brodhag/projelev/RESOURCES/index.html"
         </type>
    </DC:Relation>
</DOCUMENT>
```

This second solution enable a librarian or an expert in a particular domain to build the VWDs without the help of the authors of pages. This approach encourages a better classification of pages using ad-hoc thesauri or well-known classification schemes. The various contexts of VWDs can be stored outside the host server of pages, for example in dedicated servers, and will have to be maintained regularly by librarians.

The contexts of VWDs are either stored in a metadata.class file accessible through HTTP at the root level of WWW sites (see ALIWEB) or inside a specific domain site.

3.2 Indexing VWDs

In the following, we are interested in the indexing of the VWDs. The VWDs are represented by their context, their content which is the content of the set of pages they are built of, and their relations with other VWDs.

In the indexing phase, we keep in the index the whole context and content of VWDs, whereas only the structure links will be indexed.

In order to optimize the index traversal time, we chose to propagate the dynamic attributes along the DAG of VWDs, i.e building the index a priori. Each context is a structured information stored in a database, and the content of

a VWD is stored apart of, to process the regular expression search (for example by using the inverted-file technique). As the content of a sub-document is a part of the content of a higher-level document, we can notice that this content could be duplicated in each level of the hierarchy of the VWDs; in our solution we avoid this duplication by keeping the information relation in the index. The resulting content index associated with the root level will be the content agregation of all of its components.

3.3 Querying VWDs

The query process is based on the hierarchy of contexts and the propagation of the value of dynamic attributes for the structured part of information, combined with a textual search. The set of attributes which form together the context of a VWD are categorized into 2 types , as Fourel [8] has defined:

Definition 10 *A static attribute is local to a node, the value associated with cannot be propagated along the hierarchy of VWDs.*

Definition 11 *A dynamic attribute is ascending if its value is propagated up to the hierarchy, and it is descending if its value is propagated down to the hierarchy.*

Let us consider the following queries :
Q1 = environment
Q2 = water pollution
Q3 = water treatment + environment + pollution

In the queries Q2 and Q3, the context *environment* is explicitly specified in the user request. The usual tools would produce the following answers :
The answer to query Q1 contains many correct matches, but they are lost in a lot of noise. AltaVista gives back some 6 000 000 responses matching the word *environment*, containing pages dealing with *computers* and *ecology*. If the user tries to refine his inquery (query Q2), pages containing only *water pollution* and not *environment* are not retrieved. We have silence because the context *environment* may be implied in some relevant pages. Q3 involves silence for the same reason as Q2, whereas Q3 without *environment* induces noise, because *water treatment* occurs within both the medicine and the environment domains.

Our system supports the boolean (OR, AND, NOT) request and the regular expression over structured fields and text. Q1 is translated into ”subject:environment”, Q2 = ”subject:water pollution”, Q3 = ”subject:(environment AND pollution) AND (text:water treatment OR subject:water treatment)”. In contrast, if metadata is added to documents, the user receives the following results : Q1 returns the root documents described by the “environment” subject, instead of the whole pages containing ”environment”. Because Q2 is more detailed, the results is a set of documents speaking of “water pollution” in the environment or enterprise context.
Q3 returns the most specific page containing “treatment of water” in the recursive pollution and environment context. Looking at Fig. 3, another interesting question is Q4 = subject:(eau AND air). As the subject is both dynamic ascending and descending, the result will be the more general VWD pollution, containing subdocuments dealing both with air and water.

3.4 Experiments

We tested our search algorithm in several servers in the Ecole des Mines de Saint-Etienne. For example, the environmental site

`http://www.agora21.org`

contains more than one thousand of pages organized along a hierarchy of VWDs which can be labeled with the GEMET thesaurus. The thesaurus manage the multilinguism, associating each index term with its equivalent in other languages. The hierarchy of documents consists of sixty nodes with a depth of four levels. The main level reflects the content of the information, with the enterprise, agenda 21, Rio principles, associates, sustainable development themes.

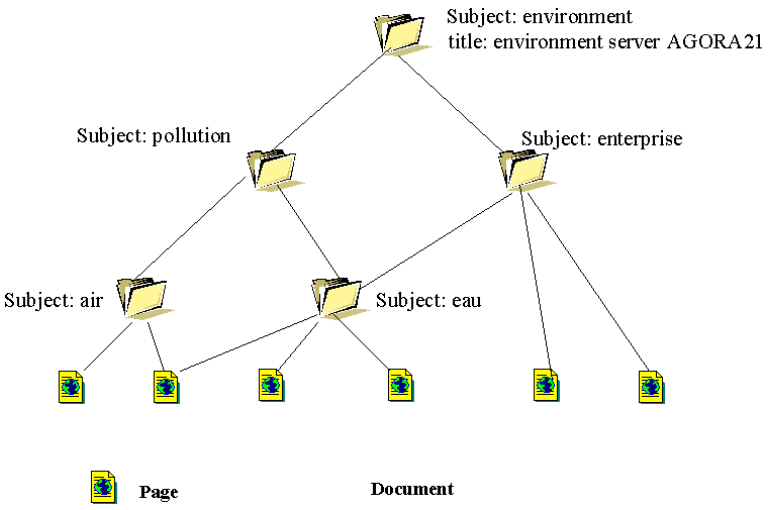

Figure 3: Hierarchy of VWDs with contexts

The whole pages have been indexed using an inverted file, the structured data reflecting the context of the VWDs are accessible from a Java application. The tree structure of the VWDs and the propagation algorithm are implemented in Java. Tools to help users to create the VWDs with contexts and to generate the relations have been implemented with the graphical Swing Java library. The Fig. 4 shows a user interface with the combined query of structured and textual information, presenting the results against the query within a hierarchy of contexts.

We can now suggest how the use of metadata associated with documents can be helpful in a cooperative architecture of search tools.

4 The model of specialized and generalized robots

4.1 Definitions

* universal robots : Universal robots currently exist on the WWW (eg. Alta Vista, Lycos), however we propose introducing a new element to universal robots to enhance their effectiveness. Although they will continue their well-known function of indexing whole pages of the WWW, they will also collect and index metadata on documents, based on the addition of a new collection files (metadata.class) which contains contexts of documents.

We define two further kinds of robots : robots for general purpose tools and robots for specialized tools.

* generalists : They are crawling the WWW to merge the metadata with their indices rather than indexing the whole URLs subtree of a WWW site. They have two functions :

  – first, they are collecting the metadata of whole sites on the WWW. They have a global view of the WWW and they index summaries of WWW documents,

  – second, they manage an acquaintance database of services (addressed by others generalists and specialists) in order to route the queries towards the right services.

* specialists : They have competencies in a particular domain. They use the metadata to decide if they are interested or not in exploring subtrees and indexing pages. They have a good algorithm for indexing information and are requested to provide final precise and consistent information.

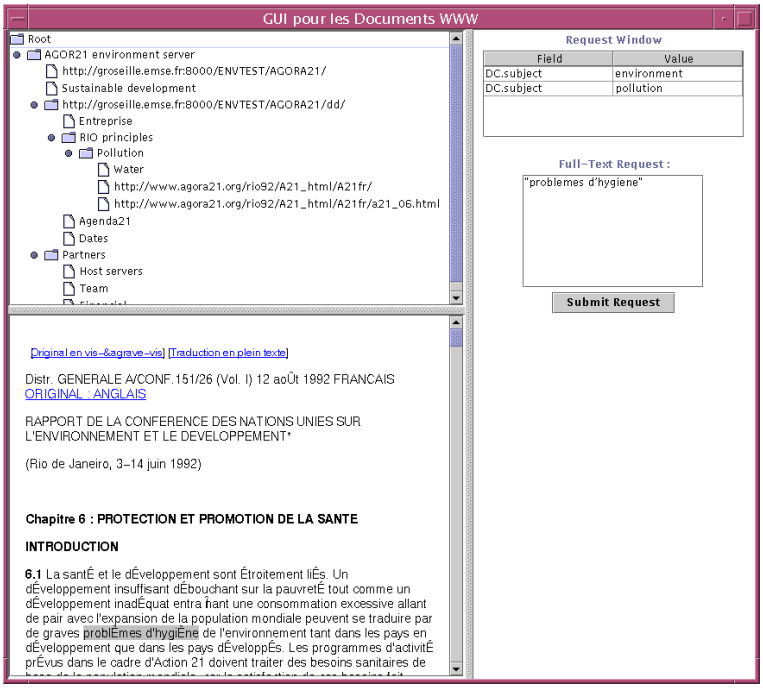

Figure 4: User Interface for querying VWDs and WWW pages in an environmental site

4.2 Architecture

4.2.1 Principles

Our architecture is based on high interactions between the entities which are participating to the model. Each specialist has competencies about its knowledge domain and can be requested by other specialists or generalists. It can describe itself thanks to metadata and give its own description to be indexed by search tools. Fig. 5 shows the design of the overall framework of generalists and specialists that cooperate using the metadata specified in the last section. In this section, we detail the function of the robots, the services they deliver and the protocols used.

The first layer is the WWW including pages and entry points to build VWDs. The second layer represents VWDs themselves and their relations described with metadata.class files stored within sites. The next layer represents the DAGs of VWDs collected by services which are universel tools, generalists or specialists tools. An administrator of VWDs can register a specialist or a generalist in order to be indexed by them or to update their knowledge database. The next layer is the architecture of cooperating tools which exchange and filter the metadata they collected from the WWW sites. The last layer is the query processor to one of the tools participating to the cooperative architecture. Three entities can outlined, as they have well-defined functions and as they surrender services.

Definition 12 *A service is represented by Service(URL, protocol, Scheme) where the URL is the HTTP access to the service, protocol enables a standardization of exchanges and Scheme specifies in which format the user or other services can access to the metadata.*

4.2.2 WWW sites providing metadata

A WWW site stores a set of WWW pages. If metadata is embedded within an existing hierarchy of documents then a site may provide :

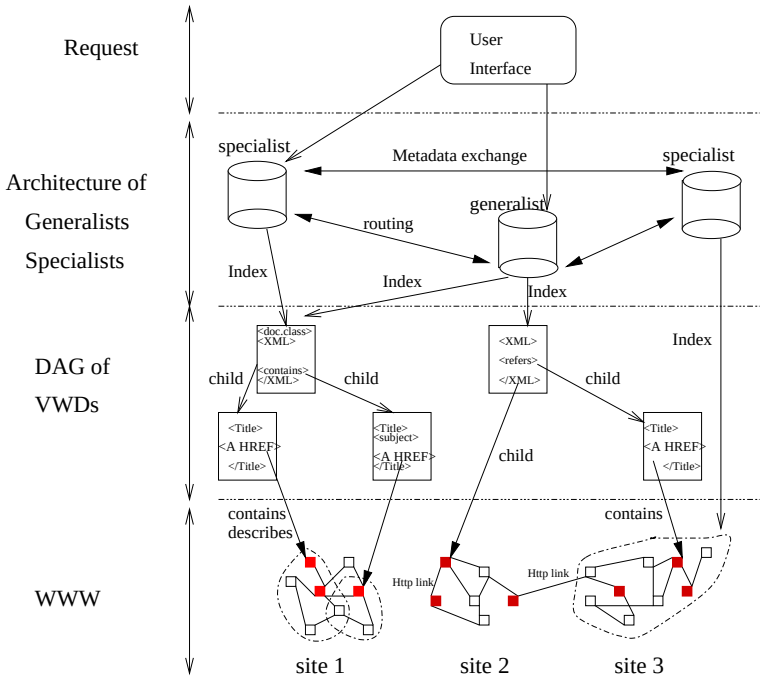

Figure 5: Architecture of the system

- one or several organizations of pages and VWDs
- a metadata.class file describing pages and VWDs
- one schema and ontology of these metadata that specify :
  - A metadata format (for instance SOIF [5], RDF [4], MCF [2])
  - One classification scheme (DDC, CDU)
  - The number of pages or volume of the site
  - Thesaurus or taxonomy if used
- a fingerprint of other services which have already indexed it.
- a standard robot.txt file

When a robot scans a site which contains metadata, it can use the metadata to decide whether to index the URLs on this site or not. It can also use this metadata instead of the documents themselves for building its indices, thus reducing the network load. In this case, it is possible to improve answers to general queries (those that give thousands of answers) : if a query generates 10000 URLs located on 100 sites, it is probably better to return the metadata associated with these 100 sites rather than a (poorly) ordered list of 10000 URLs. Such non-specific queries should be addressed to general purpose search tools.
Reciprocally, well-defined queries should be addressed to specialized search tools.

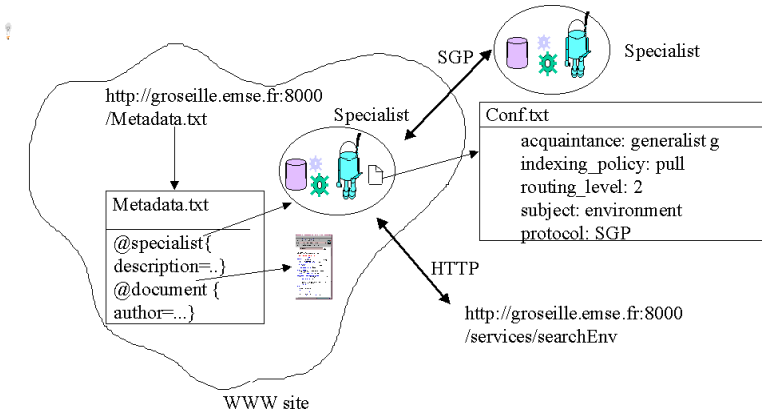

Figure 6: Implementation of specialists/generalists

4.3 Interactions between entities

4.3.1 Specialists

A specialist is created when needed and provides an identification and description of its interest domain (like environment). It registers generalists or specialists sharing the same domain, “push” information about himself. Its main function consists in gathering and indexing information concerning its domain.

- collecting pages + metadata according to some criteria (for instance VWDs belonging to an organization or a domain),
- indexing HTML pages + metadata according to some criteria,
- storing indices and metadata,
- routing requests to others specialists or generalists ,
- administrating databases (updating data, deleting invalid data…) ,
- publishing its own scheme and ontology, from a collection of various ontologies and metadata.
- keeping an acquaintance database of other tools, and a configuring file.

4.3.2 Generalists

- collecting ontologies and metadata from one or severals specialists with keeping references to the specialists
- directly collecting the metadata from the data WWW sites
- routing the refined queries to the adequate specialists or providing summarized responses to general user requests,
- adopt a political decision to collect all the metadata on the internet or not.

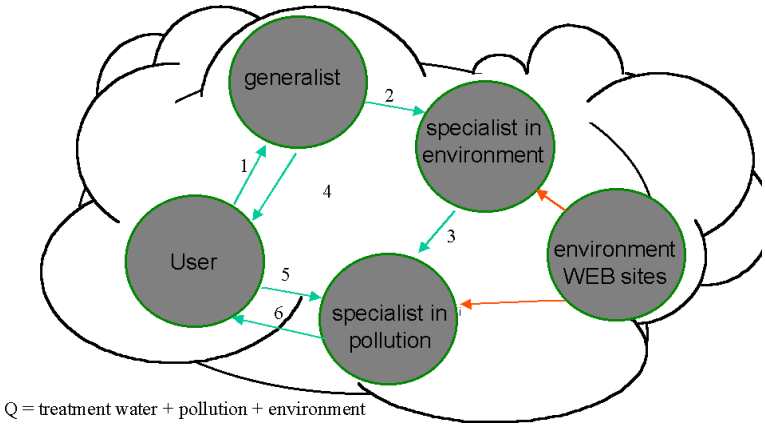

Figure 7: Routing request

4.4 Routing request

Suppose we have two specialists S1 and S2 dealing respectively with environment and pollution. Consider now the W1 site which fills S1 and S2 in environmental data, with metadata and pages. G knows S1 and its knowledge domain, whereas S1 knows S2 which is more specific than itself.

1 The user asks a generalist Q3. Q3 is translated to subject = environment, keywords = treatment + water.

2 The generalist finds one specialist in the environment area. He contacts S1 and transmits the request.

3 S1 knows another specialist, S2, whose area is specific to “pollution of water”. He transmits Q3 to S2.

4 S1 gives the user a collection of documents and pages he retrieves from his local database and shows the context of his responses with the description of S2 included.

5 S2 gives the hierarchical tree of concepts to the user.

6 The user request S2 for more detailed information.

7 S2 searches his database and gives results to the user.

5 Related work

Today, the WWW well-known search tools may be classified into two categories :
the universal search tools and the thematic directories. In this section, we study these two types of search tools and we focus particularly on the classification of WWW sites, revealing how the context and the scalability problems are dealt with current research.

5.1 Universal search tools

These search tools are universal in the sense that they try to index the whole WWW. As the WWW grows, universal tools become more numerous, resulting in overload network bandwidths and difficulties to bring the centralized index up to date. They come to be inadequate in finding relevant information all over the WWW. Main weaknesses are :

* too many irrelevant responses,

* no organization in the responses lead to difficult way to exploit them. AltaVista gives back 3 000 000 responses to the “environment” request, without explicit order,

* loss of context around the responses,

* no access to the non-textual documents (images, sounds, video) which are not easily indexed.

5.2 Classification

Organization of information is necessary for efficient information retrieval. Referring to the classifying methods coming from libraries, it is fundamental to classify documents in order to retrieve them. Classification of data has been the most useful method for organizing information in domains like library science[7] and taxonomy[15]. Standardized classification schemes emerged early in the 19th century, for example the well-known DDC (Decimal Dewey Classification). This method is called synthetical indexing, the aim is to put once in the shelves the book attached to a node of the global universal knowledge hierarchy. Another method, called analytical indexing consists on associating terms stemming from a specialized thesaurus to each document. By associating these two methods, we can describe both general and specific information and we can provide indices to retrieve them more accurately.

5.2.1 Thematic servers

We describe now the other category of available WWW search tools. As an example, Yahoo provides users with the means to browse a hierarchy of thematic directories. A provider can register by filling a form to describe its site, indicating in which topic he wishes his server to appear and at which particular level in the hierarchical tree of subjects. The advantages of this process is to enable exploratory research and better control in indexing (reducing the noise). The problems encountered by this approach are :

* manual indexing,

* only a part of WWW is indexed, so there is lot of silence in the answers,

* manual classification of the concepts and manual classification of the universal information requires a high cost for maintenance and updating,

* no content-text indexing of pages.

If we look closer to thematic servers, we can see that some classification method has been used to index the sites. Starting from a tree of concepts statically defined, each site is described by keywords that have been chosen by their author, and has a reference in the hierarchy of terms. Having Yahoo! as an example, we can either follow the ordered list of themes or ask for terms to retrieve the indexed pages.

5.2.2 Other classifications of WWW sites

The need of humans to describe data about data is necessary to perform good clustering and to improve the quality of the data delivered. That problem has been dealt with other domains, such as digital libraries, and information retrieval(IR). The Metadata Core Workshop[16] has gathered specialists in the science of information and concluded to the definition of a set of minimal metadata to describe the networked electronic information.

On the WWW environment, everybody need using metadata in order to improve the description of a collection of pages, specially on the WWW environment, because there is no means to retrieve the internal organization. To be easily indexed and retrieved, a set WWW pages need descriptions of its content and other characteristics like author, title…, then a representation of the latest thanks to metadata.

[13] provides a taxonomy and virtual URLs for browsing and searching large information spaces in an Intranet. Pan-browser[14] supports the creation, presentation and control of metadata created by users. In our model, we

allow the designer of a site to add any metadata that he feels is able to describe his documents. He can choose one classification scheme in order to disambiguate the terms.

5.3 Scalable tools

To solve the scalability problem, systems have been built upon the Internet but imply defining a new architecture (for example, Ingrid[9] has defined a new topology beyond the WWW) or propose a new hypertextual system (HyperWave[6]. Harvest[5] suggests a distributed architecture of index servers with filtering mechanisms to reduce the network bandwidth but there is not cooperation between index. HyPursuit[12] provide several co-existing hierarchies of clusters, built from an automatic clusterization of pages.

6 conclusion

To retrieve information in computer networks, research needed to define structured metadata standards embedded in the documentation to be used for organizing information and improving the construction of general indices. Here we have defined an algorithm and a propagation mechanism to improve precision and recall by expliciting the logical structure of VWDs and adding metadata to describe them. We allow the user to express contextual queries and the system gives answers embedded within different contexts and at different abstraction levels. We have a scalable architecture which offers the present search tools the ability to index quickly and with better control. Our model has the following advantages :

- Decrease consumption of the bandwidth. Robots are exchanging indices and may only index summaries of documents;
- More relevant answers. The contexts attached to the documents are hierarchically organized, involving better control upon the content of the server;
- Distributed indexing. Specialized robots are focusing the information upon one particular domain;
- A self-configuring system. Specialized and generalized robots are discovering metadata from each other.

Our future work aims to implement a hierarchic clustering algorithm to build VWD automatically, and to compare the meaningful concepts given by human with labels of concatenated words.

References

[1] Extensible markup language (xml). *http://www.w3.org/XML*.

[2] Meta content framework using xml. *http://www.w3.org/TR/NOTE-MCF-XML*.

[3] Platform for internet content selection. *The World Wide Web Consortium (W3C)*.

[4] Resource description framework (rdf). *http://www.w3.org/RDF*.

[5] C. Mic Bowman, Peter B Danzig, Darren R. Hardy, Udi Manber, and Michael F. Schwartz. The harvest information discovery and access system. *Computer Networks and ISDN Systems,28:119-125*, 1995.

[6] Wolfgang Dalitz and Gernot Heyer. *HyperWave: The New Generation Internet Information System.* 1997.

[7] A.C Foskett. *The subject approch to information.* Hamden Connecticut: Linnet Books, 1977.

[8] Franck Fourel. Impact de la structure du document sur la recherche d'information. *INFORSID, Ingénierie des systèmes d'information*, 5:339–366, 1997.

[9] Paul Francis, Takashi Kambayashi, Shin ya Sato, and Susumu Shimizu. Ingrid: A self-configuring information navigation infrastructure. *WWW 4th conference*, 1995.

[10] Massimo Marchiori. The limits of web metadata, and beyond. *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[11] C.J. Van Rijsbergen. *Information Retrieval*. 1979.

[12] Mark A. Sheldon Chanathip Namprempre Peter Szilagyi Andrej Duda David K. Gifford Ron Weiss, Bienvenido Velez. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC*, 1996.

[13] C. Fry J. Milton S. Elo, L. Weitzman. Virtual urls for browsing and searching large information spaces. *WebNet'98*, 1998.

[14] Mazer M Schickler, M and C. Brooks. Pan-browser support for annotations and other meta-information on the world wide web. *Fifth International World Wide Web Conference*, 1996.

[15] P.H.A Sneath and R.R Sokal. *Numerical Taxonomy*. 1973.

[16] Eric Miller Ron Daniel Stuart Weibel, Jean Godby. Oclc(online computer library center)/ncsa(national center for supercomputing applications) metada workshop report. *The essential elements of network object descripti on*, 1995.

Virtual WWW Documents: a Concept to Explicit the Structure of WWW Sites

Bich-Liên Doan (*doan@emse.fr)*
RIM Dpt., Ecole Nationale Superieure des Mines de Saint-Etienne
Saint-Etienne, France

Michel Beigbeder (*mbeig@emse.fr)*
RIM Dpt., Ecole Nationale Superieure des Mines de Saint-Etienne
Saint-Etienne, France

Abstract

This paper shows a new concept of a virtual WWW document (VWD), as a set of WWW pages representing a logical information space, generally dealing with one particular domain. The VWD is described using metadata in the XML syntax and will be accessed through a metadata.class file, stored at the root level of WWW sites. We'll suggest how the VWD can improve information retrieval on the WWW and reduce the network load generated by the robots. We describe a prototype implemented in JAVA, within an application in the environmental domain. The exchanges of such metadata lay in a flexible architecture based on two kinds of robots : generalists and specialists that collect and organize this metadata, in order to localize the resources on the WWW. They will contribute to the overall auto-organizing information process by exchanging their indices, therefore forwarding their knowledge each other.

Keywords: *information retrieval, WWW, search tools, metadata, cooperative architecture.*

1 Introduction

Information retrieval on the WWW has become one major research problem, partly because resources available on the WWW are heterogeneous, scattered and volatile. Another reason is the scalability problem due to the continuous growth of information volume. The WWW is viewed as a hypertext graph where the nodes are HTML pages that can be accessed through their URLs, and the links are textual anchors inside a page that point towards another URL and that enable the user to jump from page to page. HTML structures the display of pages, but provides very little information about the content and organization of pages or collections of pages; it is precisely this structured content we want.

To help users to find information on the WWW, search tools have been designed. Most of them use the inverted file technique to build indices and the boolean model to match the request with the indices of the database. The pages are indexed asindependent documents without any structure.

Besides the problem of noise and silence, the ranked list of hundred thousands of URLs given as responses is very difficult to exploit, because unless we download each HTML page, we have no idea of "is this page really tied with the subject of our question?".

In short, because the million pieces of information that build the WWW hypertext network are organized in a "flat" way, current WWW search tools are not capable of extracting the logical structure and the context around HTML pages.

To help to capture the structure and the semantics associated with each site, we propose to define the concept of virtual WWW documents as abstractions of sets of pages at different levels of granularity, which can be organized into different hierarchies of clusters in the sense of [11]. We introduce metadata that describe the VWDs precisely.

With regards to the limits of the WWW metadata, see [10], we experienced the use of metadata within a community of users, in a particular domain, the ”sustainable development”.

We propose to define a flexible architecture using the existing and available search tools on the WWW which will enable everybody to participate in the improvement of document descriptions. This architecture is based upon high interactivity between search tools and a progressive organization of information on the WWW. With minimal effort we can use cooperation between existing isolated elements of the WWW, resolving the problems of servers and bandwidth overload, scalability and context around information spaces.

In section 2 we give a model of what we call a Virtual WWW Document (VWD), we explain how to create and describe VWDs with metadata in section 3; then we describe the architecture of generalists and specialists which can take benefit from the use of metadata. Finally, we review related work and offer conclusions.

2 The Virtual WWW Document (VWD) model

We start with the following remark :
In the “printer” world, it is easy to capture the outline of a book, by reading the table of contents of the book, or its summary. On the WWW, the logical structure of pages is not explicit, even if the author of pages has a structured and hierarchical view of his information. To find out information without any ambiguity, we need to use a search word in a particular context, (for example we want to know that this set of pages is linked to environment, with environment meaning ecology and not computer environment).

In order to extract the context and organization of a set of HTML pages, we propose to define virtual WWW documents, we give a representation of them and we use metadata to describe them.

The metadata are composed of both pieces of information about the content of the document and external to that content. For instance, at the book level the author name, the publication date and the title are outside its semantical content.

Our system is based on two simultaneous level access : the data level (content of pages) and the metadata level (set of attribute-value pairs describing the context of pages).

2.1 Definitions

In this section we define the concepts and relations manipulated by the system : resource, VWD, relation, context, ontology, scheme, site.

Definition 1 *A HTML page is a WWW resource which is identified and accessed through its URL. The relation Page(URL, title, date, text) contains the title, last modification date, text.*

Definition 2 *A hypertext link is represented by a relation Link(source, label, target) where source is the URL of the resource the link start from, label is the text bound to the link and target is the URL of the destination page.*

We give now a definition of metadata :

Definition 3 *Litterally “data about data”, metadata is machine understandable information about WWW resources or other things. Metadata has a semantics and a structure about people, things, concepts or ideas.*

Definition 4 *An ontology is a domain specific metadata : ”objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them”.*
We define Ontology(name, ref), where name is a String identifying the ontology, for example ”thesaurus GEMET” in the environmental domain, ref is the URL of the service or database providing the vocabulary and relations between terms.

Definition 5 *A scheme is a data model used to represent the data. To describe the information in our system, we specify the set of elements of metadata, the syntax and the semantics of each element. Scheme(name, metadata) is identified by its name, a set of elements of metadata, for example the Meta Data Dublin Core.*

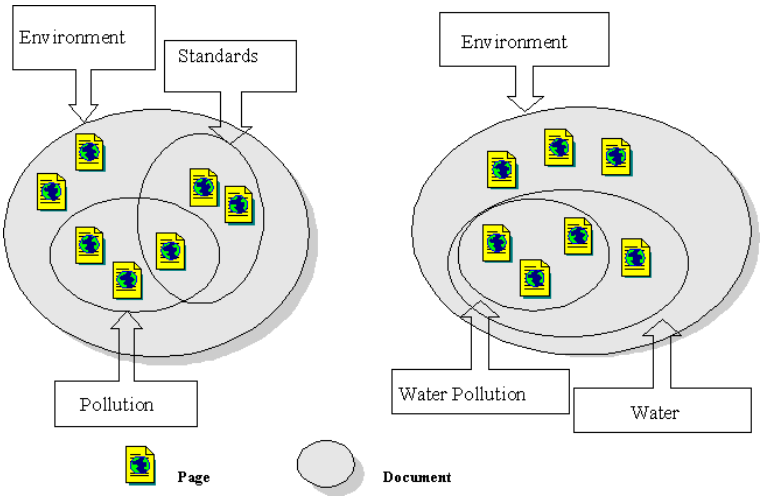

Figure 1: Several organizations of the same set of environmental pages

This scheme may be stored within the site, at the root place or may be a reference to a URL stored somewhere else.

Definition 6 *A relation is represented by Relation(name, type) where name is the name of the N:M relationship between VWDs, or VWD and HTML pages. Type is the type of the relationship.*

Examples of relations are association, child, contains, reference .... We use two important relations :
- child is the composition relationship with other subdocuments;
- contains is the set of HTML pages belonging relationship.

We introduce now the notion of context around a page or a document : Each page or document belongs to a context. Let's go back to the library example, and consider a page of a book. This page belongs to a section, this section belongs to a chapter which belongs to the book itself. Each level of organization of the book gives a context, from the highest level (the root is the book) to the next level in the hierarchy up to the page level. These contexts may inherit from their father context, depending if the attributes are dynamic descending (can be propagated down to the hierarchy). So we suggest to make explicit these contexts within the WWW documents, by giving some information about pages or set of pages contained within a document. The Fig. 1 shows how the WWW pages can be organized into clusters called VWDs, one page can belong to different documents, therefore bounded to different contexts.

Definition 7 *A context is represented by a tuple Context(name, Ontology, Scheme, $\{Field\}$), where name is a string uniquely identifying the context, a set of 4-tuple Field(attribute, qualifier, value, type) where :*

- *attribute is the name of the field;*
- *qualifier brings in more precision to the field;*
- *value is a set of keywords belonging to an Ontology or not;*
- *the type of value $\in \{String, Int, Float, Boolean, Date\}$.*

Definition 8 *Let us give a definition of a VWD :*
a collection of pages clustered according to some criteria, for instance pages dealing with the same subject like environment, representing a semantic information unit which can be given as a response to a user's request.
A VWD is represented by VWD(URL, Context, [Relation]) where :

- *URL is the identifier of the page or service describing the VWD;*
- *Context is the metadata linked to the VWD;*

- Relation is a typed link between VWDs, for example the link child enables to define a Directed Acyclic Graph (DAG) of VWDs.

Each page can belong to one or several documents. A document itself can be a component of other documents.

Definition 9 *A WWW site is characterized by Site(domain, {resource },[{VWD }], [metadata.class]) where :*

- domain is the domain name of the server hosting the WWW resources (for example www.w3c.org);
- resource is any data available on a WWW site through a URL;
- VWD is a collection of WWW pages or resources which can be explicited and organized into a DAG of documents;
- metadata.class file is the context of the VWDs available on a WWW site, and stored at the root level of the site, accessible through a URL.

The last two components of this relation are optional, we see later how they can be generated and maintained.

2.2 Describing a VWD with metadata

This semantics and structure may be different, according to the creator of metadata, and it would be naive to consider that only one standardized metadata will be used by every one. The recommendations given by PICS[3], RDF[4], including a meta-metadata to define the type of descriptions used will be adopted, in conjunction with the semantics chosen by OCLC/NCSA Workshop[16].

In the following, we give an example of our metadata. Our model requires the exchange of indices between different entities, in an extensible and flexible way. XML is a document description language [1], which enables the representation of structured document, and which is flexible and simple for the authors to use. Moreover, it will be translated by the next versions of "browsers" from Netscape and Microsoft.

So we opt for a XML syntax and we include the semantics of Metadata Dublin Core.

```
<?xml version="1.0"?>
<?xml:namespace ns="http://groseille.emse.fr/DC" Prefix="DC"?>
<!DOCTYPE DOCUMENT "http://www.emse.fr/~brodhag/projelev#d0.">
<DOCUMENT>
    <DC:Language>" fr-French" </DC.Language>
    <DC:Title> "environment" </DC:Title>
    <DC:Subject>
        <DC:scheme>
                <DDC> " 333.3" </DDC>
        </DCscheme>
    </DC:Subject>
    <DC:Subject> "pollution" </DC:Subject>
    <DC:Relation>
         <type="child">
                "http://www.emse.fr/~brodhag/projelev/RESOURCES"
         </type>
    </DC:Relation>
</DOCUMENT>
```

3 Creating, indexing and querying the VWDs

In this section, we introduce the problem of creating the VWDs, how these documents are maintained and by whom, finally we explain how to retrieve them.

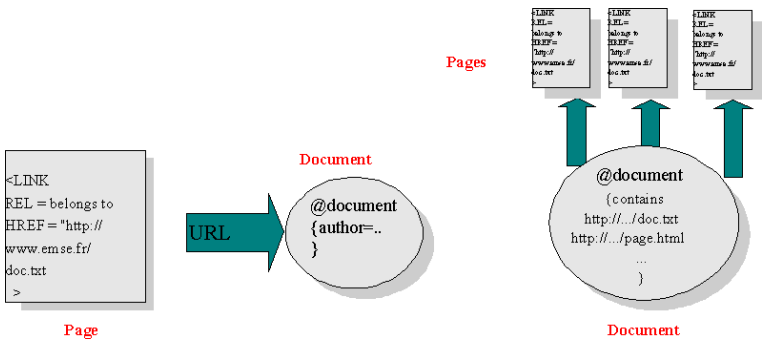

Figure 2: Creation of VWDs

3.1 Creation of VWDs and their context

Two solutions can be adopted to generate the VWDs. First the author of HTML pages add META tags inside its pages to indicate that these pages belong to one (or several) VWD located and described in a metadata.class file in the host server. For example he uses

```
<META NAME="containedIn"
CONTENT="http://groseille.emse.fr/metadata.class#d0>
```

This solution is practical for the administration of VWDs because if a page is deleted or modified, no change has to be done in the VWDs. If author are encouraged to put META keywords, description, subject inside the pages, then it will help to apply clustering algorithm to automatically build the hierarchies of VWDs.

A second consists in creating only the VWDs and using the relation contains to include the set of pages inside the VWD.

```
<DOCUMENT>
  <DC:Relation>
         <type="contains">
         "http://www.emse.fr/~brodhag/projelev/RESOURCES/index.html"
         </type>
    </DC:Relation>
</DOCUMENT>
```

This second solution enable a librarian or an expert in a particular domain to build the VWDs without the help of the authors of pages. This approach encourages a better classification of pages using ad-hoc thesauri or well-known classification schemes. The various contexts of VWDs can be stored outside the host server of pages, for example in dedicated servers, and will have to be maintained regularly by librarians.

The contexts of VWDs are either stored in a metadata.class file accessible through HTTP at the root level of WWW sites (see ALIWEB) or inside a specific domain site.

3.2 Indexing VWDs

In the following, we are interested in the indexing of the VWDs. The VWDs are represented by their context, their content which is the content of the set of pages they are built of, and their relations with other VWDs.

In the indexing phase, we keep in the index the whole context and content of VWDs, whereas only the structure links will be indexed.

In order to optimize the index traversal time, we chose to propagate the dynamic attributes along the DAG of VWDs, i.e building the index a priori. Each context is a structured information stored in a database, and the content of

a VWD is stored apart of, to process the regular expression search (for example by using the inverted-file technique). As the content of a sub-document is a part of the content of a higher-level document, we can notice that this content could be duplicated in each level of the hierarchy of the VWDs; in our solution we avoid this duplication by keeping the information relation in the index. The resulting content index associated with the root level will be the content agregation of all of its components.

3.3 Querying VWDs

The query process is based on the hierarchy of contexts and the propagation of the value of dynamic attributes for the structured part of information, combined with a textual search. The set of attributes which form together the context of a VWD are categorized into 2 types , as Fourel [8] has defined:

Definition 10 *A static attribute is local to a node, the value associated with cannot be propagated along the hierarchy of VWDs.*

Definition 11 *A dynamic attribute is ascending if its value is propagated up to the hierarchy, and it is descending if its value is propagated down to the hierarchy.*

Let us consider the following queries :
Q1 = environment
Q2 = water pollution
Q3 = water treatment + environment + pollution

In the queries Q2 and Q3, the context *environment* is explicitly specified in the user request. The usual tools would produce the following answers :
The answer to query Q1 contains many correct matches, but they are lost in a lot of noise. AltaVista gives back some 6 000 000 responses matching the word *environment*, containing pages dealing with *computers* and *ecology*. If the user tries to refine his inquery (query Q2), pages containing only *water pollution* and not *environment* are not retrieved. We have silence because the context *environment* may be implied in some relevant pages. Q3 involves silence for the same reason as Q2, whereas Q3 without *environment* induces noise, because *water treatment* occurs within both the medicine and the environment domains.

Our system supports the boolean (OR, AND, NOT) request and the regular expression over structured fields and text. Q1 is translated into ”subject:environment”, Q2 = ”subject:water pollution”, Q3 = ”subject:(environment AND pollution) AND (text:water treatment OR subject:water treatment)”. In contrast, if metadata is added to documents, the user receives the following results : Q1 returns the root documents described by the “environment” subject, instead of the whole pages containing ”environment”. Because Q2 is more detailed, the results is a set of documents speaking of “water pollution” in the environment or enterprise context.
Q3 returns the most specific page containing “treatment of water” in the recursive pollution and environment context. Looking at Fig. 3, another interesting question is Q4 = subject:(eau AND air). As the subject is both dynamic ascending and descending, the result will be the more general VWD pollution, containing subdocuments dealing both with air and water.

3.4 Experiments

We tested our search algorithm in several servers in the Ecole des Mines de Saint-Etienne. For example, the environmental site

`http://www.agora21.org`

contains more than one thousand of pages organized along a hierarchy of VWDs which can be labeled with the GEMET thesaurus. The thesaurus manage the multilinguism, associating each index term with its equivalent in other languages. The hierarchy of documents consists of sixty nodes with a depth of four levels. The main level reflects the content of the information, with the enterprise, agenda 21, Rio principles, associates, sustainable development themes.

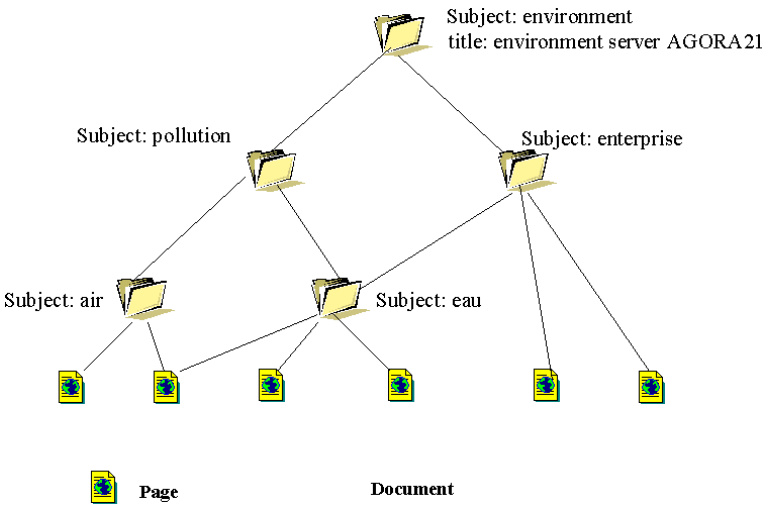

Figure 3: Hierarchy of VWDs with contexts

The whole pages have been indexed using an inverted file, the structured data reflecting the context of the VWDs are accessible from a Java application. The tree structure of the VWDs and the propagation algorithm are implemented in Java. Tools to help users to create the VWDs with contexts and to generate the relations have been implemented with the graphical Swing Java library. The Fig. 4 shows a user interface with the combined query of structured and textual information, presenting the results against the query within a hierarchy of contexts.

We can now suggest how the use of metadata associated with documents can be helpful in a cooperative architecture of search tools.

4 The model of specialized and generalized robots

4.1 Definitions

* universal robots : Universal robots currently exist on the WWW (eg. Alta Vista, Lycos), however we propose introducing a new element to universal robots to enhance their effectiveness. Although they will continue their well-known function of indexing whole pages of the WWW, they will also collect and index metadata on documents, based on the addition of a new collection files (metadata.class) which contains contexts of documents.

We define two further kinds of robots : robots for general purpose tools and robots for specialized tools.

* generalists : They are crawling the WWW to merge the metadata with their indices rather than indexing the whole URLs subtree of a WWW site. They have two functions :

  – first, they are collecting the metadata of whole sites on the WWW. They have a global view of the WWW and they index summaries of WWW documents,

  – second, they manage an acquaintance database of services (addressed by others generalists and specialists) in order to route the queries towards the right services.

* specialists : They have competencies in a particular domain. They use the metadata to decide if they are interested or not in exploring subtrees and indexing pages. They have a good algorithm for indexing information and are requested to provide final precise and consistent information.

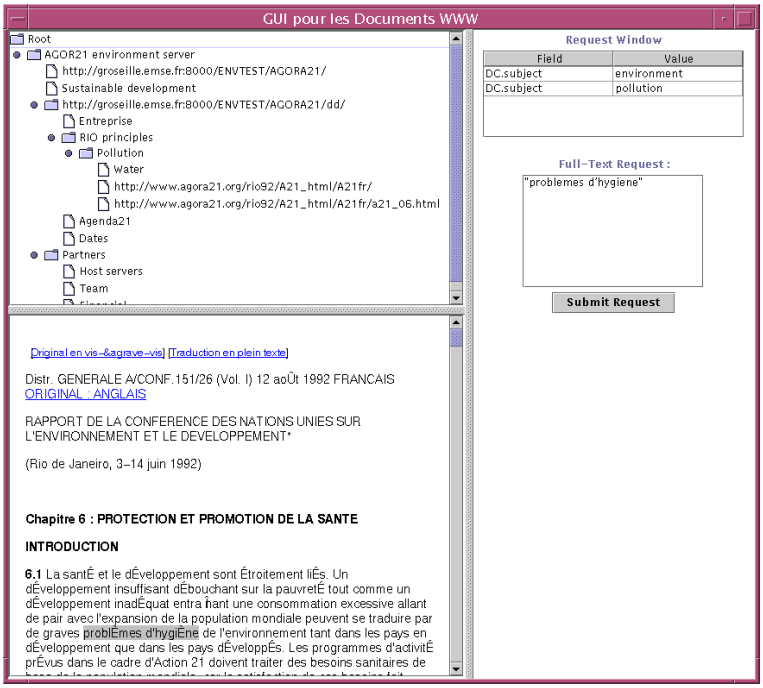

Figure 4: User Interface for querying VWDs and WWW pages in an environmental site

4.2 Architecture

4.2.1 Principles

Our architecture is based on high interactions between the entities which are participating to the model. Each specialist has competencies about its knowledge domain and can be requested by other specialists or generalists. It can describe itself thanks to metadata and give its own description to be indexed by search tools. Fig. 5 shows the design of the overall framework of generalists and specialists that cooperate using the metadata specified in the last section. In this section, we detail the function of the robots, the services they deliver and the protocols used.

The first layer is the WWW including pages and entry points to build VWDs. The second layer represents VWDs themselves and their relations described with metadata.class files stored within sites. The next layer represents the DAGs of VWDs collected by services which are universel tools, generalists or specialists tools. An administrator of VWDs can register a specialist or a generalist in order to be indexed by them or to update their knowledge database. The next layer is the architecture of cooperating tools which exchange and filter the metadata they collected from the WWW sites. The last layer is the query processor to one of the tools participating to the cooperative architecture. Three entities can outlined, as they have well-defined functions and as they surrender services.

Definition 12 *A service is represented by Service(URL, protocol, Scheme) where the URL is the HTTP access to the service, protocol enables a standardization of exchanges and Scheme specifies in which format the user or other services can access to the metadata.*

4.2.2 WWW sites providing metadata

A WWW site stores a set of WWW pages. If metadata is embedded within an existing hierarchy of documents then a site may provide :

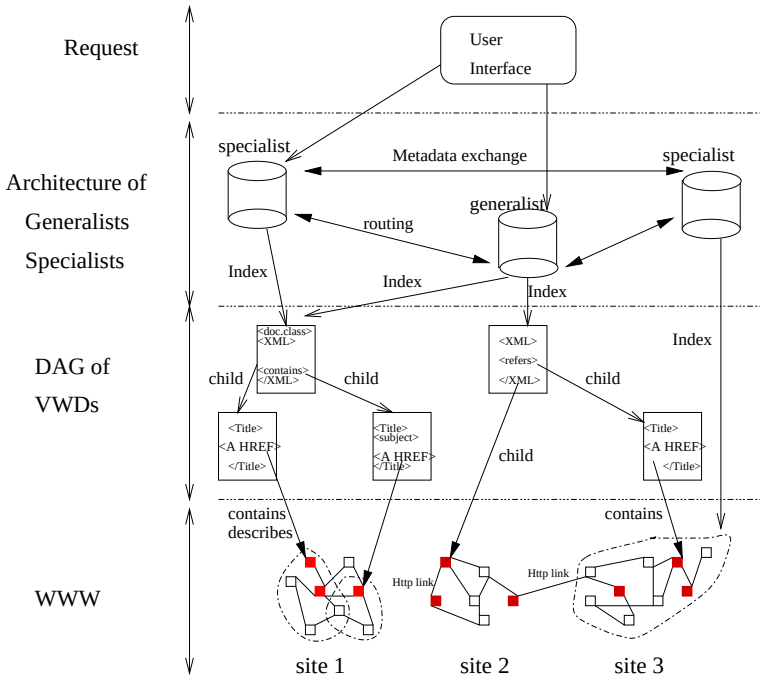

Figure 5: Architecture of the system

- one or several organizations of pages and VWDs
- a metadata.class file describing pages and VWDs
- one schema and ontology of these metadata that specify :
  - A metadata format (for instance SOIF [5], RDF [4], MCF [2])
  - One classification scheme (DDC, CDU)
  - The number of pages or volume of the site
  - Thesaurus or taxonomy if used
- a fingerprint of other services which have already indexed it.
- a standard robot.txt file

When a robot scans a site which contains metadata, it can use the metadata to decide whether to index the URLs on this site or not. It can also use this metadata instead of the documents themselves for building its indices, thus reducing the network load. In this case, it is possible to improve answers to general queries (those that give thousands of answers) : if a query generates 10000 URLs located on 100 sites, it is probably better to return the metadata associated with these 100 sites rather than a (poorly) ordered list of 10000 URLs. Such non-specific queries should be addressed to general purpose search tools.
Reciprocally, well-defined queries should be addressed to specialized search tools.

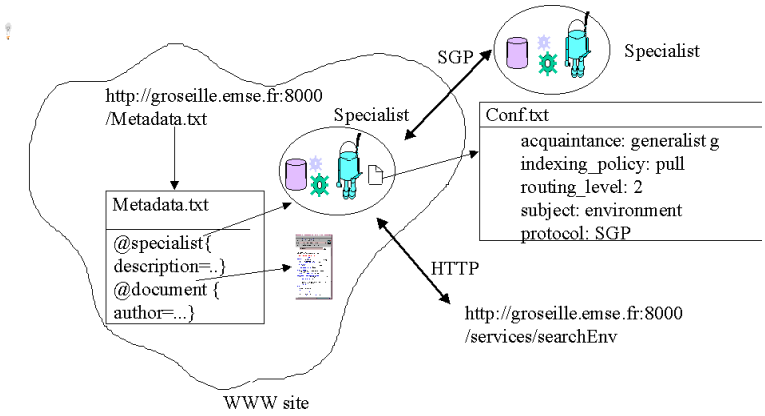

Figure 6: Implementation of specialists/generalists

4.3 Interactions between entities

4.3.1 Specialists

A specialist is created when needed and provides an identification and description of its interest domain (like environment). It registers generalists or specialists sharing the same domain, “push” information about himself. Its main function consists in gathering and indexing information concerning its domain.

- collecting pages + metadata according to some criteria (for instance VWDs belonging to an organization or a domain),
- indexing HTML pages + metadata according to some criteria,
- storing indices and metadata,
- routing requests to others specialists or generalists ,
- administrating databases (updating data, deleting invalid data…) ,
- publishing its own scheme and ontology, from a collection of various ontologies and metadata.
- keeping an acquaintance database of other tools, and a configuring file.

4.3.2 Generalists

- collecting ontologies and metadata from one or severals specialists with keeping references to the specialists
- directly collecting the metadata from the data WWW sites
- routing the refined queries to the adequate specialists or providing summarized responses to general user requests,
- adopt a political decision to collect all the metadata on the internet or not.

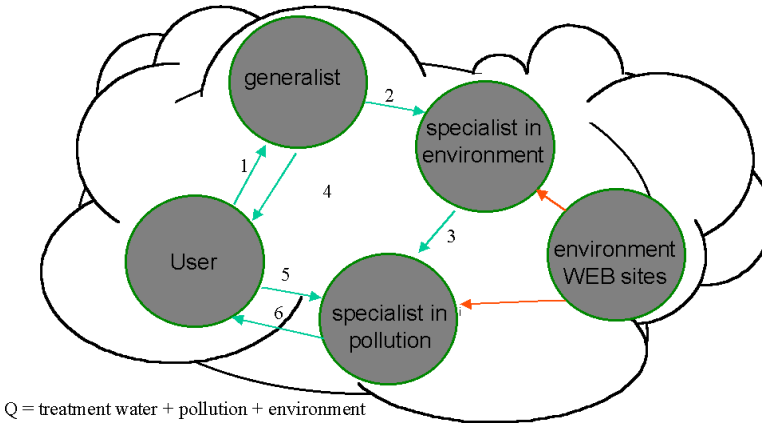

Figure 7: Routing request

4.4 Routing request

Suppose we have two specialists S1 and S2 dealing respectively with environment and pollution. Consider now the W1 site which fills S1 and S2 in environmental data, with metadata and pages. G knows S1 and its knowledge domain, whereas S1 knows S2 which is more specific than itself.

1 The user asks a generalist Q3. Q3 is translated to subject = environment, keywords = treatment + water.

2 The generalist finds one specialist in the environment area. He contacts S1 and transmits the request.

3 S1 knows another specialist, S2, whose area is specific to “pollution of water”. He transmits Q3 to S2.

4 S1 gives the user a collection of documents and pages he retrieves from his local database and shows the context of his responses with the description of S2 included.

5 S2 gives the hierarchical tree of concepts to the user.

6 The user request S2 for more detailed information.

7 S2 searches his database and gives results to the user.

5 Related work

Today, the WWW well-known search tools may be classified into two categories :
the universal search tools and the thematic directories. In this section, we study these two types of search tools and we focus particularly on the classification of WWW sites, revealing how the context and the scalability problems are dealt with current research.

5.1 Universal search tools

These search tools are universal in the sense that they try to index the whole WWW. As the WWW grows, universal tools become more numerous, resulting in overload network bandwidths and difficulties to bring the centralized index up to date. They come to be inadequate in finding relevant information all over the WWW. Main weaknesses are :

* too many irrelevant responses,

* no organization in the responses lead to difficult way to exploit them. AltaVista gives back 3 000 000 responses to the “environment” request, without explicit order,

* loss of context around the responses,

* no access to the non-textual documents (images, sounds, video) which are not easily indexed.

5.2 Classification

Organization of information is necessary for efficient information retrieval. Referring to the classifying methods coming from libraries, it is fundamental to classify documents in order to retrieve them. Classification of data has been the most useful method for organizing information in domains like library science[7] and taxonomy[15]. Standardized classification schemes emerged early in the 19th century, for example the well-known DDC (Decimal Dewey Classification). This method is called synthetical indexing, the aim is to put once in the shelves the book attached to a node of the global universal knowledge hierarchy. Another method, called analytical indexing consists on associating terms stemming from a specialized thesaurus to each document. By associating these two methods, we can describe both general and specific information and we can provide indices to retrieve them more accurately.

5.2.1 Thematic servers

We describe now the other category of available WWW search tools. As an example, Yahoo provides users with the means to browse a hierarchy of thematic directories. A provider can register by filling a form to describe its site, indicating in which topic he wishes his server to appear and at which particular level in the hierarchical tree of subjects. The advantages of this process is to enable exploratory research and better control in indexing (reducing the noise). The problems encountered by this approach are :

* manual indexing,

* only a part of WWW is indexed, so there is lot of silence in the answers,

* manual classification of the concepts and manual classification of the universal information requires a high cost for maintenance and updating,

* no content-text indexing of pages.

If we look closer to thematic servers, we can see that some classification method has been used to index the sites. Starting from a tree of concepts statically defined, each site is described by keywords that have been chosen by their author, and has a reference in the hierarchy of terms. Having Yahoo! as an example, we can either follow the ordered list of themes or ask for terms to retrieve the indexed pages.

5.2.2 Other classifications of WWW sites

The need of humans to describe data about data is necessary to perform good clustering and to improve the quality of the data delivered. That problem has been dealt with other domains, such as digital libraries, and information retrieval(IR). The Metadata Core Workshop[16] has gathered specialists in the science of information and concluded to the definition of a set of minimal metadata to describe the networked electronic information.

On the WWW environment, everybody need using metadata in order to improve the description of a collection of pages, specially on the WWW environment, because there is no means to retrieve the internal organization. To be easily indexed and retrieved, a set WWW pages need descriptions of its content and other characteristics like author, title…, then a representation of the latest thanks to metadata.

[13] provides a taxonomy and virtual URLs for browsing and searching large information spaces in an Intranet. Pan-browser[14] supports the creation, presentation and control of metadata created by users. In our model, we

allow the designer of a site to add any metadata that he feels is able to describe his documents. He can choose one classification scheme in order to disambiguate the terms.

5.3 Scalable tools

To solve the scalability problem, systems have been built upon the Internet but imply defining a new architecture (for example, Ingrid[9] has defined a new topology beyond the WWW) or propose a new hypertextual system (HyperWave[6]. Harvest[5] suggests a distributed architecture of index servers with filtering mechanisms to reduce the network bandwidth but there is not cooperation between index. HyPursuit[12] provide several co-existing hierarchies of clusters, built from an automatic clusterization of pages.

6 conclusion

To retrieve information in computer networks, research needed to define structured metadata standards embedded in the documentation to be used for organizing information and improving the construction of general indices. Here we have defined an algorithm and a propagation mechanism to improve precision and recall by expliciting the logical structure of VWDs and adding metadata to describe them. We allow the user to express contextual queries and the system gives answers embedded within different contexts and at different abstraction levels. We have a scalable architecture which offers the present search tools the ability to index quickly and with better control. Our model has the following advantages :

- Decrease consumption of the bandwidth. Robots are exchanging indices and may only index summaries of documents;
- More relevant answers. The contexts attached to the documents are hierarchically organized, involving better control upon the content of the server;
- Distributed indexing. Specialized robots are focusing the information upon one particular domain;
- A self-configuring system. Specialized and generalized robots are discovering metadata from each other.

Our future work aims to implement a hierarchic clustering algorithm to build VWD automatically, and to compare the meaningful concepts given by human with labels of concatenated words.

References

[1] Extensible markup language (xml). *http://www.w3.org/XML*.

[2] Meta content framework using xml. *http://www.w3.org/TR/NOTE-MCF-XML*.

[3] Platform for internet content selection. *The World Wide Web Consortium (W3C)*.

[4] Resource description framework (rdf). *http://www.w3.org/RDF*.

[5] C. Mic Bowman, Peter B Danzig, Darren R. Hardy, Udi Manber, and Michael F. Schwartz. The harvest information discovery and access system. *Computer Networks and ISDN Systems,28:119-125*, 1995.

[6] Wolfgang Dalitz and Gernot Heyer. *HyperWave: The New Generation Internet Information System.* 1997.

[7] A.C Foskett. *The subject approch to information.* Hamden Connecticut: Linnet Books, 1977.

[8] Franck Fourel. Impact de la structure du document sur la recherche d'information. *INFORSID, Ingénierie des systèmes d'information*, 5:339–366, 1997.

[9] Paul Francis, Takashi Kambayashi, Shin ya Sato, and Susumu Shimizu. Ingrid: A self-configuring information navigation infrastructure. *WWW 4th conference*, 1995.

[10] Massimo Marchiori. The limits of web metadata, and beyond. *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[11] C.J. Van Rijsbergen. *Information Retrieval*. 1979.

[12] Mark A. Sheldon Chanathip Namprempre Peter Szilagyi Andrej Duda David K. Gifford Ron Weiss, Bienvenido Velez. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC*, 1996.

[13] C. Fry J. Milton S. Elo, L. Weitzman. Virtual urls for browsing and searching large information spaces. *WebNet'98*, 1998.

[14] Mazer M Schickler, M and C. Brooks. Pan-browser support for annotations and other meta-information on the world wide web. *Fifth International World Wide Web Conference*, 1996.

[15] P.H.A Sneath and R.R Sokal. *Numerical Taxonomy*. 1973.

[16] Eric Miller Ron Daniel Stuart Weibel, Jean Godby. Oclc(online computer library center)/ncsa(national center for supercomputing applications) metada workshop report. *The essential elements of network object descripti on*, 1995.