Semantic Analysis of a Web site: A prototype

Michel Sala
LIRMM
16 rue Ada 34392
Montpellier, France
sala@lirmm.fr

Isoird Gael
LIRMM
16 rue Ada 34392
Montpellier, France
isoird@lirmm.fr

Abstract

Many works has been made to improve the users browsing on Web sites notably with the adaptation techniques which allow to guide the users according to their profile in order to direct them to the most interesting pages. Considerable improvements has been made in this domain but it is very difficult to judge the impact of the implementation of such techniques on a site. In this paper, we are going to present a prototype which we set up to study the architecture of a Web site with regard to its semantics. The aim is to try to judge the Web site modelling quality by studying the users browsing with regard to various presented concepts. It allows to have an effective tool to advance the defects of certain Web sites in the optics to improve presentation to the users.

Introduction

In this paper, we are going to present a prototype which we conceived to try to judge the quality of a Web site without leaning on statistical data but on site semantic[1].
Necessaries elements for the project functioning are:

- The Web site graph representing the set of pages and links existing among them
- The Web site ontology regroup in a hierarchy organized way the different concepts which it approaches
- The server connection logs of the Web site (called logs files) allowing the users categorization [2][3]

We suppose that the studied sites use formalism XML. If it is not case, we suppose that there is a preliminary phase allowing to translate sites into XML.[7]
The main objective of this work is to have an effective tool to consider the conception quality of a Web site with regard to its semantics. After the use of such a tool, the sites designers shall be able to see which are various problems such as the little visited pages or those on that the users do not stay. So, they will be able to reshape the site to improve the frequence of their site and the users navigation.

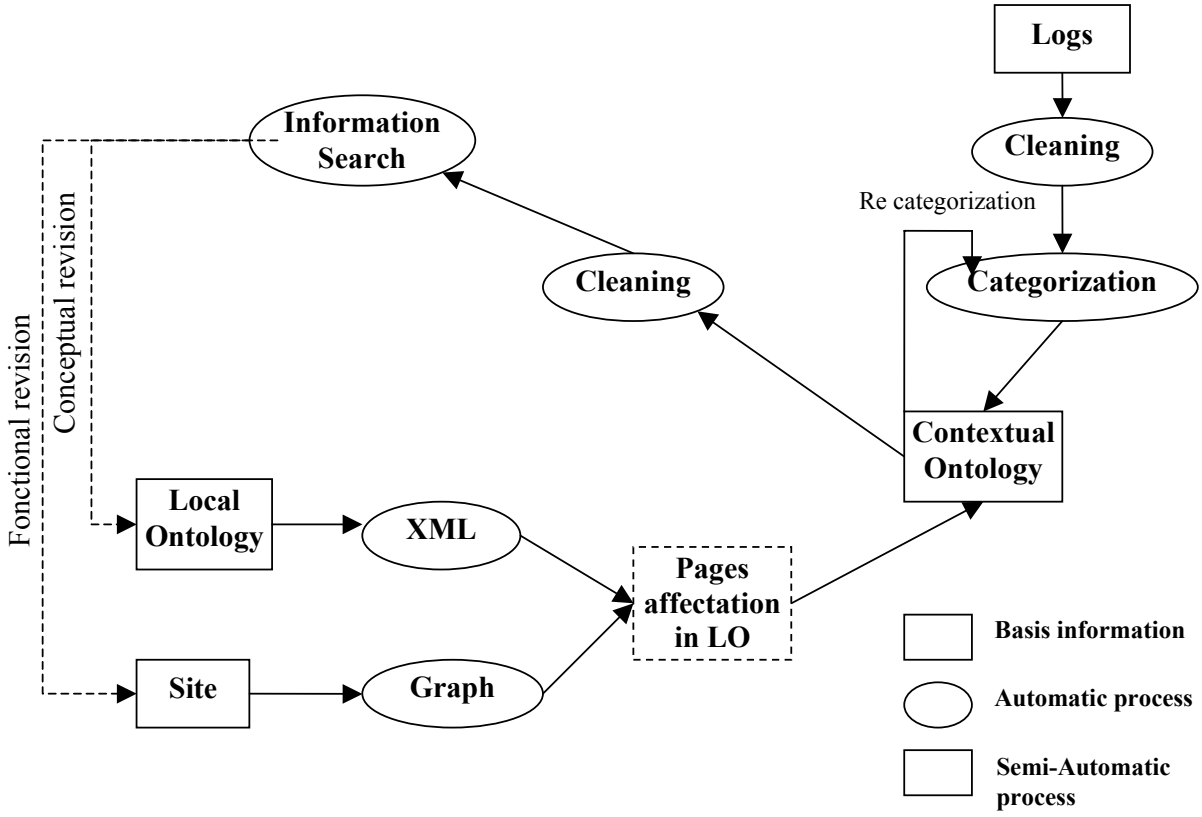
Presentation of the prototype architecture

In this paragraph, we are going to see the various stages which allow to make a semantic study of a Web site.
Our project can be structured in five very different phases which are :

- creation of the site graph
- construction of the site ontology
- categorization of the users

- use of the site reference ontology to extract specific subontologies with some users' categories
- exploitation of ontologies extracted to discover information which allow to judge the quality of the site.

General progress of our project can be resume by the following figure.



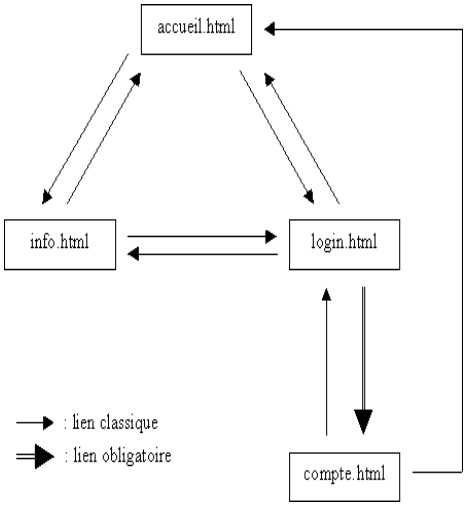
General Process

We are now going to see the most important steps of the semantic study process on a Web site.

The creation of the site graph:

To be able to draw a user browsing on a Web site, it is necessary to represent the dynamics of this one, that is the set of its pages and links connecting them.
This dynamics can be easily represented under shape of a directed graph with values in which, nodes would be the pages of the site, the arcs hypertext links between the different pages, and boolean arcs values indicating if link is compulsory, that is if the page destination can be hited only having visited the page of origin.
To store the site graph so defined, we decided to use a simple text file in order to facilitate the manual reading and publishing phase. This file will contain two parts, the first listing the various site pages (graph nodes), and the second different links between these pages (graph arcs). These links are represented with the following formalism: [({has, b}, c); d] knowing that a indicates the URL of the origin page, b the URL of the destination page, c the boolean

(T for True, and F for False) indicating if link is compulsory and d a supplementary information for the link (a priori, not used).




```
http://www.exemple.fr/accueil.html
http://www.exemple.fr/info.html
http://www.exemple.fr/login.html
http://www.exemple.fr/compte.html


[({http://www.exemple.fr/accueil.html,http://www.exemple.fr/info.html},F);nc]
[({http://www.exemple.fr/accueil.html,http://www.exemple.fr/login.html},F);nc]
[({http://www.exemple.fr/info.html,http://www.exemple.fr/accueil.html},F);nc]
[({http://www.exemple.fr/login.html,http://www.exemple.fr/accueil.html},F);nc]
[({http://www.exemple.fr/compte.html,http://www.exemple.fr/accueil.html},F);nc]
[({http://www.exemple.fr/info.html,http://www.exemple.fr/login.html},F);nc]
[({http://www.exemple.fr/login.html,http://www.exemple.fr/info.html},F);nc]
[({http://www.exemple.fr/login.html,http://www.exemple.fr/compte.html},V);nc]
[({http://www.exemple.fr/compte.html,http://www.exemple.fr/login.html},F);nc]
```

Example of the graph site and representation

The first phase of our project consists so in creating the text file representing the studied site graph, from the site plan or directly the Web site.

The creation of the site ontology

Beyond the functional knowledge of the site described by the graph, it is necessary, for our project, to know the site semantics, and more exactly that of each pages [5]. To answer this requirement, it will be necessary to use a precise model. The ontological graph model is the most pratical because it allows, not only to list the set of concepts approached on a site, but also to express the links which can be among them (these links clarifying a specialization or a generalization of the considered concept).
For our project, it was chosen to approach description by ontology in a very simple way. Consequently, rather than to use a structure of ontological graph, we decided to use ontological arborescences. So, the concepts of the studied site will be treated on a hierarchical basis, but these links of hierarchical organization will not carry semantic indication (relation of specialization, composition), in order to not compex the model.
An ontology so built will be called a reference ontology for the concerned site, or still local ontology (LO) of the site, by opposition to global ontologies (GO) describing they domains in its entirety, regrouping several LO of sites making left for these domains. The creation of the local ontology for a site supposes the existence of XML files and XMLSchéma describing it.

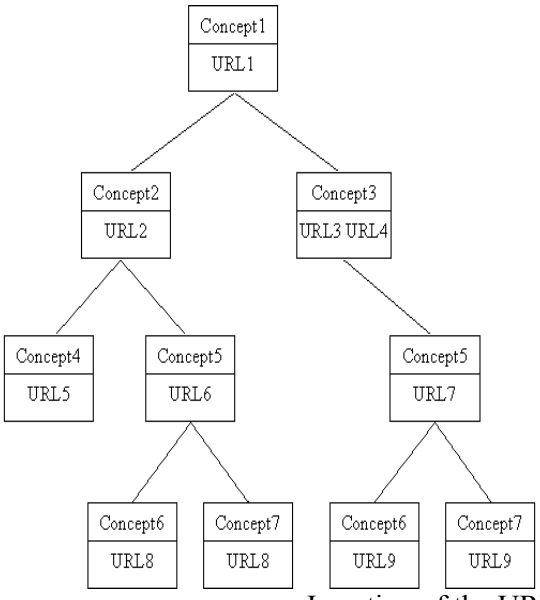
```
<xs:element name="Concept1">
   <xs:complexType>
      <xs:sequence>
         <xs:element name="Concept2">
            <xs:complexType>
               <xs:sequence>
                  <xs:element name="Concept4" type="xs:string"/>
                  <xs:element name="Concept5" type="Concept5Type"/>
               </xs:sequence>
            </xs:complexType>
         </xs:element>
         <xs:element name="Concept3">
            <xs:complexType>
               <xs:sequence>
                  <xs:element name="Concept5" type="Concept5Type"/>
               </xs:sequence>
            </xs:complexType>
         </xs:element>
      </xs:sequence>
   </xs:complexType>
</xs:element>

<xs:complexType name="Concept5Type">
   <xs:sequence>
      <xs:element name="Concept6" type="xs:string"/>
      <xs:element name="Concept7" type="xs:string"/>
   </xs:sequence>
</xs:complexType>
```

```
<concept name="Concept1">
   <concept name="Concept2">
      <concept name="Concept4"/>
      <concept name="Concept5">
         <concept
name="Concept6"/>
         <concept
name="Concept7"/>
      </concept>
   </concept>
   <concept name="Concept3">
      <concept name="Concept5">
         <concept
name="Concept6"/>
         <concept
name="Concept7"/>
      </concept>
   </concept>
</concept>
```

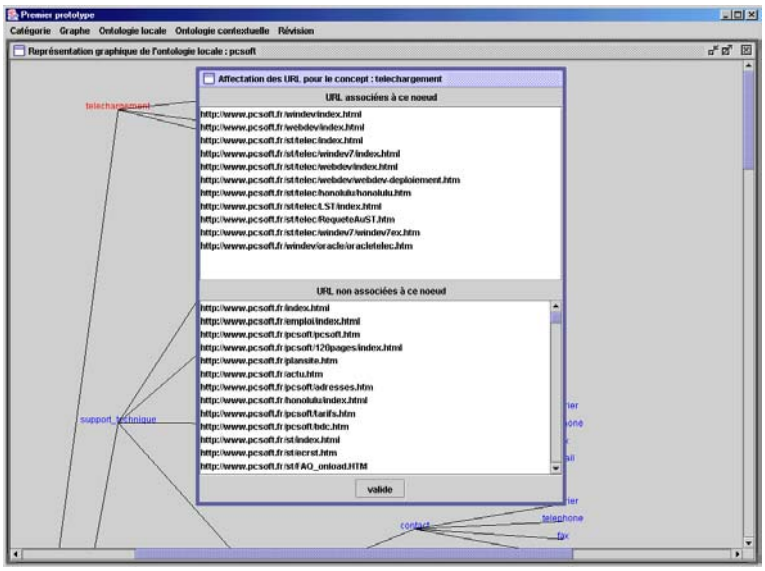
XMLSchéma to the local ontology

When this local ontology was created, we uses the site graph to specify which are the pages which treat every concept. This stage is important because it is due to this new ontology called labelled local ontology that we shall be able to make a correlation between the users browsing and the visited concepts.




```
<concept name="Concept1">
   <URL name="URL1"/>
   <concept name="Concept2">
      <URL name="URL2"/>
      <concept name="Concept4">
         <URL name="URL5"/>
      </concept>
      <concept name="Concept5">
         <URL name="URL6"/>
         <concept name="Concept6">
            <URL name="URL8"/>
         </concept>
         <concept name="Concept7">
            <URL name="URL8"/>
         </concept>
      </concept>
   </concept> ..........
```

Insertion of the URL in the local ontology



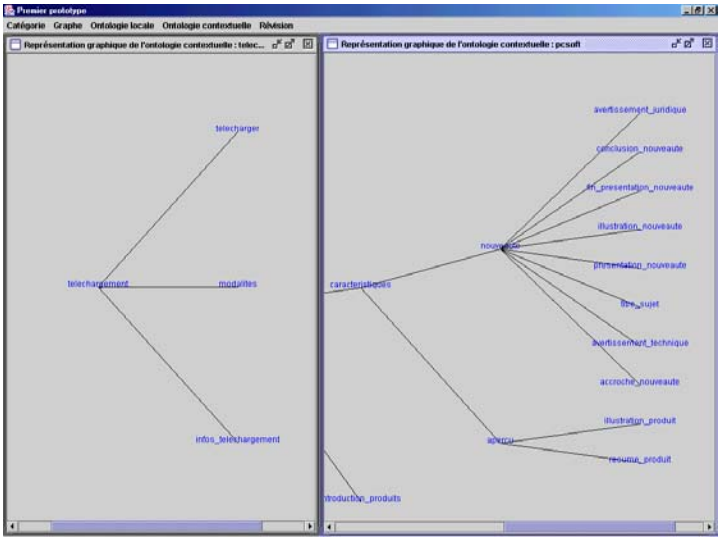
Linking the ontologie and the graph site

Users categorization

To realize the usersactivity on a site, it is necessary to study the conexions logs. To be exploitable and significant, the contents of this logs will be reformatted under shape of users' categories according to statistical criteria. Every category will be established by users having a similar behavior by crossing the site. To obtain this categorization, logs will be first of all "cleaned", to keep only necessary information (access only to the site pages) and relevant (pages visited during a significant time), which will be then regrouped by user, then by category.

Creation of contextual ontologies :

Having established local ontology labelled with the site and its users' categories, and in order to put in common these two elements to extract information of it, it is necessary to build for every users' category a contextual ontology, that is an ontology appropriate for this category. A contextual ontology (CO) is an ontology extracted from the local ontology of the studied site. It is conceived so as to be a relevant element of information. It is at the same an ontology, and so present a result under an unified, coherent and complete shape within the framework of the domain described by the local ontology.
Indeed, every users category comparable to a set (or browsing) of visited pages, we can associate to these pages, due to the local ontology, a sub-domain (i.e. a set of concepts) of the local ontology. So, obtained contextual ontologies will describe, in a sense, the centre of interest of every users' category, to the studied Web site.



Example of contextual ontology

Information search :

For a given site, its logs will allow to determine certain number of users' categories, from which we shall deduct so many vast categories. It is by means of these categories and CO corresponding that we shall try to put in evidence the possible structural defects at the navigation level of the site.We can find more probably possible ergonomics defects (by statistics on differences between vast categories and users' associated categories) being able to be owed, for example , to little visible links. The discovery of such defects would pulla revision called functional, because applying to the site (and its functions). On the other hand, this search stage for information being also able to reveal possible defects on the local ontology of the site, it is necessary also to foresee another type of revision called abstract, because applying to the concepts structure approached on the site.

Tests and results :

We made a work in association with the site of PC soft [8], what allowed us to test our prototype and to supply information about the site conception. Obtained results were very interesting and allowed to give advices to the site designer in order to improve various points. First of all, we discovered pages which not referenced so inaccessible for the Internet users. Then, we studied the semantic coherence of customers browsing and noticed some incoherence due to a bad organization of the concepts chain. Several remarks ensued from our analysis and recommended to the site designers to make two types of revisions, a functional (re-modelling of the structure of the site) and the other one abstract aiming to remove semantic incoherence discovered in the customer browsing. These two revisions had several results. At the level of the site and of its structure, one obtains a more homogeneous syntactic set. One tries to reduce the site depth by creating links to shorten navigation. At the customers level, navigation is definitly improved so much in term of time saving but also in term of semantic coherence. Groups of users are more easily created with propositions of personalized browsing.

Conclusion

In this paper, we presented the prototype which we set up to realize the semantic study of a Web site. For it we detailedthe various analysis phases of our prototype. We base ourselves on the graph of the site, on its ontology and on the logs files study. From this information we look has to create groups of users following their behavior and to put in correlation these groups with the ontology containing the site pages for every concept. So, we can see the groups centres of interest and make a search for information aiming to find which are concepts and pages which are not well to frequent. This information will be useful for the site designer who will be able to reshape it to improve frequence and navigation of the users.

Perspectives

Due to our prototype, we could improve the conception of all the sites, notably those where the study of the Internet user browsing is important. We think well on various commercial stakes, but also improvement of services supplied with Web as for example on-line forming or E-learning. The semantic analysis of Web sites lets also suspect the possibility of making semantic comparisons of sites what shall allow to have criteria allowing to define if a site relating to a particular domain is more or less complete that another site speaking about the same domain. But above all this prototype aims to improve theInternet users navigation [4] on site and the most important perspectives stay those of the proposition of personalized browsing whether it is by user classification from its arrival with regard to its necessities or by management adaptative browsing

References

[1]. Sala, D. Hérin, P. Pompidor : Aid to the Semantic Maintenance of the Web Site Proceedings of the Conference on Sciences Electroniques, Technologies de l'Information et des Télécommunications (SETIT), Tunisie, march 2004

[2] R. Srikant, R. Agrawal, Mining Sequential Patterns : Generazations and Performance Improvements. In Proceedings of the 5th Internatiopnal Conference on Extending Database Technology (EDBT'96), p 3-17, Avignon, France Sept 96

[3] F. Masseglia, P. Poncelet, M. Teisseire : Using Data Mining Techniques on Web Acces Logs to Dynamically Improve Hypertext Structure

[4] P. Brusilovsky : Adaptive Hypermedia and Attempt to Analyze and Generalize. In Workshop on Adaptive hypertext and hypermedia at UM'94, Hyannis, Cape Cod, MA, USA.

[5] M-S Segret, P. Pompidor, D. Hérin, M. Sala : Use of ontologies to integrate some information semi-structured exists of pages web, INFORSID'2000, Lyon pp 37-55

[6] D. Hérin, M. Sala, P. Pompidor : Evaluating and revising browsing from web ressources educationnal, ITS 2002, Springer-Verlag LNCS, pp 208-218, june 2002

[7] www.w3.org/XML

[8] PcSoft Web Site www.pcsoft.fr