Towards automatic XML structure building for Web documents

Agnès Guerraz

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Copy title from file/Properties/Summary/Title

Insert author here

N° ????

Août 2003

THÈME 1

ISSN 0249-6399

Towards automatic XML structure building for Web documents

Agnès Guerraz[1]

Thème 1 –Format and notation, Markup languages
Projet Wam

Rapport de recherche n° ???? – 28/02/07 - 17 pages

Abstract: Web documents avalaible through the Internet are frequently supplied simply as poorly-written HTML or as plain text. Indeed, almost all of these Web documents are understandable only by humans, staying unexploitable by softwares and computers. The power of Semantic Web tools and XML technologies can only be deployed on documents having a minimum of formalism in their structure. This paper relates to the structuration process for Web documents that do not have a real structure through markup languages such as XML or definition of grammars for validing them. It deals with building of structure in documents when existing structure is insuficient or inexistant. This subject is closely related to the problems of automatic creation of XML schemas or templates. This work lies concretely within the scope of XML documents and their problems, related to the fact that their structure building and set up is time consuming for the user. Being based on techniques of data mining, information of structures is captured, clarifying and returning the names and the characteristics of structure elements, in particular their relationships, their constraints and their logical organization. This paper proposes a process which makes it possible to calculate automatically elements of structures (1) by applying methods of data mining on documents, (2) by building components of structure automatically, (3) by automatically proposing XML transformations on the final structured document. Initially, this work will use all the range of schemas going from XML schemas to templates.

Keywords: Web, XML, adaptation, data mining, schema, template.

[1] WAM – Agnes.Guerraz@inria.fr

Unité de recherche INRIA Rhône-Alpes
655, Avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 – Télécopie : +33 4 76 61 52 52

Towards automatic XML structure building for Web documents

Résumé: Mon projet de recherche concerne la structuration de contenu de documents web pas ou peu structurés. Il s'agit de construire la structure de documents web qui n'en ont pas au départ ou dont la structure est insuffisante. Cette notion est étroitement liée à la problématique de création automatique de schémas XML ou de templates. Ce travail s'inscrit concrètement dans le cadre des documents XML et de la problématique de création de documents structurés. Grâce à des techniques de data mining, les informations de structures sont capturées rendant explicite le sens des noms des éléments, et exploitant certaines caractéristiques des éléments notamment leurs liens, contraintes et leur organisation logique. Ce projet propose un processus qui permet de calculer automatiquement des éléments de structures (1) en appliquant des méthodes de data mining sur le(s) documents web, (2) en construisant automatiquement des composants de structure, (3) en proposant automatiquement des transformations XML sur le document final structuré. Ce travail utilisera toute la gamme de schémas allant des schémas XML aux templates.

Mots clés: web, XML, adaptation, data mining, Schéma XML, template.

1 Introduction

The Web contains a huge among of heterogenous data which are structured, semi-structured, textual, multimedia (sound, images, videos,...) or even mathematical formulas. These data should at the end make sense to humans who are carrying them and also to computers and machines using them inside their softwares.

The fields of analysis of this research work are the Web and XML technologies. It particularly involves two fields of research: the Semantic Web and the Web mining which develop quickly and build both on the success of the Web [24]. Each one addresses a share of a new challenge posed by the current Web: the nature of the majority of the data on the Web is so much little or badly structured that they can only be understood by humans, but the quantity of data is so enormous that they can be treated eficiently only by computers. The need for structuring Web documents correctly and easily seems to be a challenge.

This research report treats of this problem, rather vast, and which will solve applicative problems. It is a question of creating more Semantic Web by exploiting the ”hidden data” automatically extracted from current Web pages.

The purpose of this research report is to present a process for the automatic building of Web document structures. In the first part, the global situation of the scientific context of this concern is presented, then in the second part our approach is shown specifying which methods and which algorithms to use, in the third part solutions are exposed to achieve this process, and following these three parts, a discussion shows how this process could help in enhancing our current Web.

2 Situation

Information available on the Internet is frequently provided in a form which is not exploitable for data processing specialists who intend these data for their softwares or their information system. A significant part of the Web documents is available in a kind of not exploitable HTML which will be transformed into rough text in order to clarify the data and to remove this HTML which can induce in error.

Thus, these documents are like simple texts the structure of which is made by orthography and semantics conventions [20], comprehensible for the humans but not for softwares. The need for structuring the documents seems here obvious to make it possible for the computers to process these documents and to carry out intelligent operations. XML was created to exchange a large variety of structured documents as well as data on the Web [10]. Although it is possible to send XML on the Web and to leave the treatment of the format for the presentation to the client side, in the majority of cases, XML is used only on the server side, as a source format from which the other representations are declined.

Documents are transformed into XHTML before being delivered to the client side. As this treatment is tiresome - to define a diagram or a DTD, to convert the document into XHTML, while often using XSLT transformations - many documents available on the Web are very little or even not structured [11].

The Web raises questions of scale, multi-media data, and temporal information. The users of this information as well as the suppliers are facing problems related to this very nature of the Web [22].
For the user, the problems are to find adequate and correct information, to create new knowledge from those on the Web, to personalize available information according to her/his preferences of contents and presentation. For the supplier, it is a question of understanding and of learning what users want, and of customization for each user and each use.

2.1 Semantic Web

The semantic Web takes more and more of importance at the research level. The best known definition is the one given by Tim Berners-Lee in an article [9] published in Scientific American in May 2001:
"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

That means that information does not have a well defined significance in the current Web. It is indeed true from a point of view of data processing specialist since the major part of information is in textual form, very little structured, and thus unusable to perform treatments of calculation or inferences. It is however quite obvious that information available on the current Web has a significance, but which it is accessible today only to human readers. The vision of the Semantic Web, in which information will be accessible and easy to handle automatically by computers and machines, can be summarized finally by a pile of languages (figure 1) playing each one a particular role:

- XML provides a manner of representing structured documents;
- XML Schema makes it possible to enforce the structure of XML documents;
- RDF is a simple data model, based on resources and relationships between these resources, equipped with Semantics and which can be represented in XML;
- RDF Schema makes it possible to define the vocabulary to describe classes and properties, hierarchically in taxonomies;

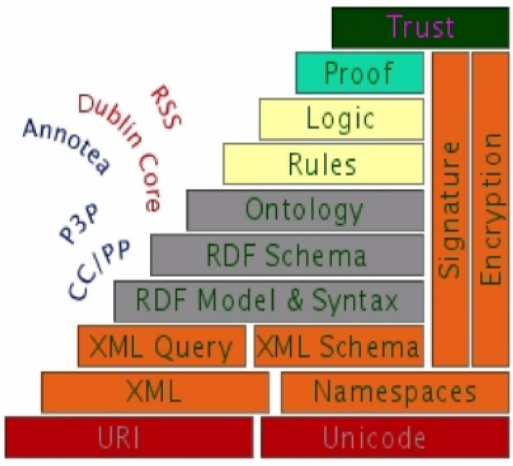

figure 1. The Layer Cake - from Tim Berners Lee.

2.2 Web Mining

Web Mining [13] aims at extracting and exploiting useful knowledge coming from the Web. This research domain consists in using the whole of Data Mining technologies in order to develop approaches and tools, making it possible to extract relevant information starting from data of the Web (documents, interaction traces, page structures, links...).

The Web Mining is divided into three approaches, the Web Structure Mining, the Content Web Mining and the Web Usage Mining. The first relates to the structural analysis of the Web, the second to the study of the contents of Web pages and finally the third to the study of the behaviors of user navigation. More particularly, the Web Content Mining is interested in the contents of the Web pages. To this end, the techniques of description, classification and analysis of terms are very useful to treat the textual part of the pages. The Web Content Mining is also interested in images, and especially in the links inside Web pages, to process discovered sources of information through the Web pages. For example, it makes it possible for each Web page to quantify the images, the text zones...
By the joint analysis of the page hit ratio (Web Usage Mining), it is possible to determine if pages which contain more images (more attractive but longer to load) are more visited than pages containing more text.

The question which interests us is to know how linguistic technologies will help to structure new information which is and will be published on the Web in the next years.

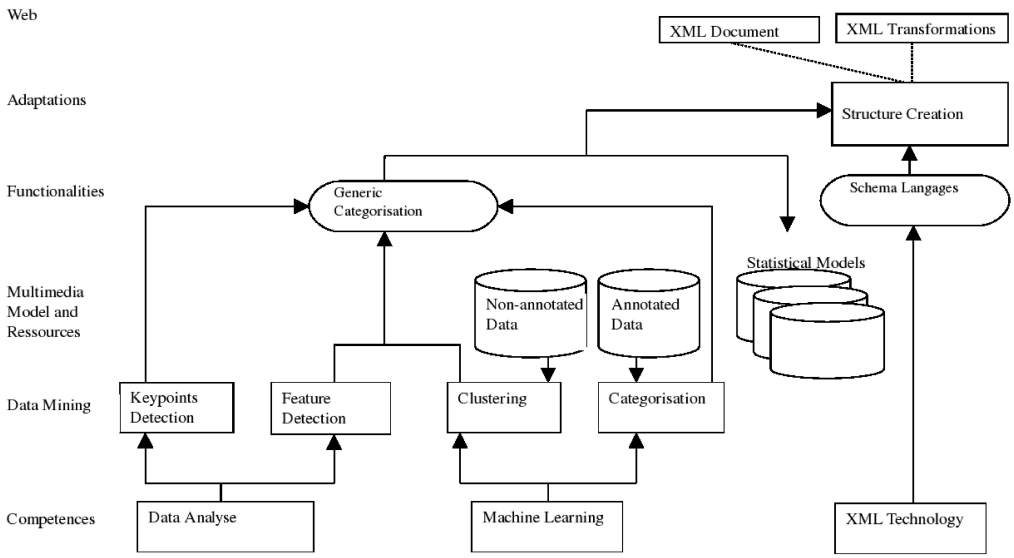

figure 2. Procedure of creation of document structure.

3 Methods

The selected analysis methods and the particular approach for this research work are described on what was made in Data Mining; this work falls under the tradition of the automatic treatment of languages. The automatic categorization of documents is made possible thanks to supervised training. It aims at providing a data-processing program that is able to assign, in an autonomous way, documents to their category. A training program is carried out on a whole document collection to which category labels were already assigned by the humans. The

clustering is an equivalent process but it is a non-supervised one, and it requires no training and thus no labelling work.

This research work aims at setting up an automatic treatment for building, for extracting one or more basic structures of template type or in a better case: of XML schema, which are based on the data resulting from the statistical model. For any Web document of the starting collection there will be one (or several) XML schema or template. While being always based on the statistical model to which the document belongs, this research also seeks to exploit the statistical data of the model to create automatic transformations of the document: for example to extract information (XMLQuery) by questioning the document compared to the keywords of its category.

Following the machine learning approach, a number of algorithms specifically designed for automatic markup have been described in the literature [8].
MarkitUp! [16] is an early predecessor of automatic markup systems. It is a training of the structure of semi-structured documents (e.g. HTML) starting from examples in order to make a SGML document from it. The method is based on the grammar generation by generalization (unification of derived rules provided by the user).
SRV [23] considers all possible phrases as potential slot fillers. A multistrategy approach combines evidence from three classifiers (rote learner, naiveBayes classifier, relational rule learner).

3.1 Data Mining

Data Mining makes it possible to carry out compact and comprehensible models returning account of relations binding the description of the document collection to a result relating to this description.
The essential difference between traditional statistics and Data Mining is that the techniques of Data Mining build the aforementioned model in an automatic way whereas the traditional statistical techniques require to be handled and guided by a professional statistician. The techniques of Data Mining make it possible to build automatically a model of dependences on data.

We will use Data Mining, categorization and clustering, to extract from semantic concepts and to use them to produce XML schema and transformations, as shown in the figure 2.
This approach consists in studying the statistical model of multimedia documents that are more or less structured, and to build in an interactive way the statistical model while being based on bayesian operators. The innovation of this approach lies in the automatic construction of structure elements thanks to the exploitation of the built statistical model.

3.2 Input Document Codage

In this direction, documents are transcribed in a vectorial format that is named Bag Of Word (BOW). The units will be the tokens for textual information, the chunks for audio information, the KeyPoints for the images (figure 2). Indeed, data treated by this project can be of multimedia nature. It appears that the Hidden Markov Model (HMM) constitutes nowadays the most usually used and powerful technique for the extraction of information. The HMM will be used at this step.

The HMM is a statistical method which models sequences of states, named hidden states as they are non-observable. The model includes probabilities of transition between these states and the probabilities of emission starting from these states in order to model the observations.

Either q_t the state of the system and y_t the output at t, the emission of each state $P(y_t|q_t)$ is probabilist and depends only on the current state q_t. The transition between two states $P(q_t|q_{t-1})$ is also probabilist and depends only on the preceding state. The HMM, in fact, is based on homogeneous Markov chains since the dynamics of the system is only given by the probabilities of transition which are time independent.
In order to effectively take advantage of the HMM, it is necessary to impose a topology on the state graph. The aim of this topology is to obtain a better control on the number of free parameters and to be able to inject a knowledge on the nature of the data. A topology is characterized by the presence or the absence of transitions between states. Once topology chosen, the training of the HMM can be carried out by using the algorithm called Expectation-Maximization (EM) [12].

The HMM are generative models. That implies that for a task of classification, a distinct HMM must be involved for each class considered. A Naive Bayes (NB) classifier can be used, with prior probabilities equal for each component. The result will be gathered inside a data structure of Bag Of Keypoints (BOK). Then, at this step, the whole of the data is tagged or marked out thanks to the HMM.

After that, comes the step of building the data model. There are several possible methods according to the collected data:

- the supervised mode - the data are annotated manually according to a well-known list of classes or categories in order to build the model. It is the most immediate and easy case;
- the unsupervised mode - the data are not annotated and the algorithm gathers the data according to a number of classes or categories pre-estimated by the user and, thanks to specific refinement operators of the result, we make it converged towards a final model;
- the semi-supervised mode contains at the same time annotated and not annotated data.

When the model is built we can finally carry out the categorization or classification of BOK.

3.3 Statistical Model Building

An important step will consist in building the statistical model of the data (figure 2). The Naive Bayes model constitutes a reference in the field of the training. NB is a probabilistic approach resting on the theorem of the conditional probabilities of Bayes. NB model rests on the assumption of independence of the attributes relatively to their class.
In the case of textual documents indexed by words, it correspond to the formula 1.

FORMULA 1.

$$P(d) = \sum_{\alpha} P(\alpha)P(d/\alpha) = \sum_{\alpha} P(\alpha) \prod_{w \in V} P(w/\alpha)^{(d,w)}$$

where *d* represents a document, *w* a word of the document collection vocabulary *V*, *#(d, w)* the number of occurrences of the word *w* in the document *d* and where α described the set of the latent classes.
NB works strictly with labelled data (classified data). Its effectiveness strongly depends on the cardinal of the set of training documents. At the end of the training, the algorithm is able to predict, with a certain relevance, the class of a non labelled document.

There are many probabilistic models used: a review of the various models which one gathers under the denomination of Naive Bayes is made in [14] and how it is possible to convert these methods into clustering algorithms [18].

The algorithms of classification of documents have to treat an increasingly important volume of information and labelling is a long, tiresome, realizable task only by the humans [1] when it is indeed possible. Building tools allowing the automation of this stage of categorization is crucial [18] to enable us to navigate and to control the increasing mass of documents. This project brings a new dimension by using these algorithms to the end of structuring web documents from the point of view of the semantic Web. A certain number of models will be examined, certain requirements will have to be satisfied and to be based on the criteria given in [17], as summarized here:

- some **polythematicity** of the documents, and thus to carry out a multi-assignment of the documents to the classes and categories;
- some **polysemia** and **synonymy** of the attributes (we hear here these terms in a broad direction, generalizing with the other attributes the properties of the words). These models must thus be able on the one hand to assign the same attribute to several classes, and on the other hand to gather different attributes from close "directions";
- some **structure** of a whole of classes or categories. These models must thus be able to structure the classes in the case of classification, and to exploit an existing structure in the case of categorization.

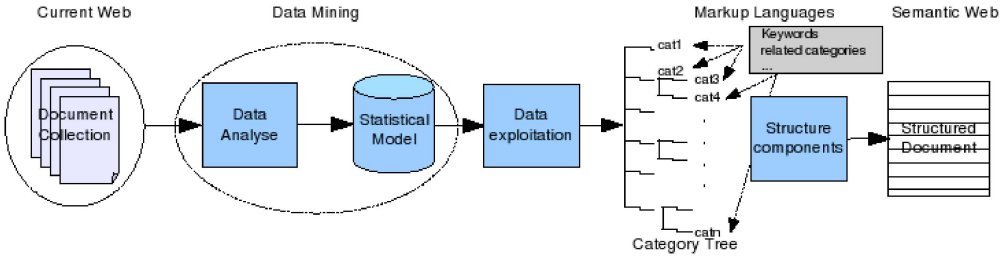

figure 3. Extraction of the document structure starting from the document collection.

4 Solutions

As shown previously in this research report, this research concern aims at better exploiting the Web document collections, and at using from the extracted information to provide exploitable structured documents in order to carry out intelligent XML transformations on documents which at the beginning of the process were poorly or non structured. Research tasks on the automatic conversion on a semi-structured and heterogeneous document corpus towards a preset XML schema already showed good results [24] while working on the search for information in texts using the processing of semi-structured data (XML), the classification and the clustering of structured documents to make re-structuration of documents. Our concern is further interested in the documents that are not or poorly structured, to build structure elements, even the whole structures when it is possible.

In their article [8], Abolhassani and al, give a survey on existing approaches. They roughly distinguish between three types of markup:

- Macro-level markup deals with the global visual and logical structure of a document (e.g. part, chapter, section, down to the paragraph level.)
- Micro-level markup is used for marking single words or word groups. For example, in news, person and company names, locations and dates may be marked up, possibly along with their roles in the event described (e.g. a company merger).
- Symbol-level markup uses symbolic names as content of specific elements in order to describe content that is not plain text (e.g. MathML for mathematical formulas and CML for chemical formulas).

Compared to their approach, our resarch focuses on the development of a markup which is not related to a specific domain. Our automatic building of document structure will aim at extracting automatically entities and their relationships such as names, people functions, location names, organisation names, but also more semantic concepts such as professional experience or education levels in the example of a document collection containing resumes.

The innovation of our process is to build the document schema according to a statistical model (figure 3). Depending on this statistical model, there will be in a range of structure descriptions which can go from XML schema to template. On the basis of Web document collection, we apply Data Mining algorithms allowing to build the statistical model of this Web document collection within the meaning of the criteria given in [17] . At the end of this process, we obtain a tree of n categories which have a keyword list for each category, vectors of close categories and indicators on distances and their probabilities. It is possible to apply this process to smaller units, founded on a logical cutting of the document, like the paragraph units in the case of textual units, which will be treated like a document inside the above process.

Then, to create the schema of the document, various approaches will be evaluated according to the obtained data.

The structure of XML documents are normally described using a language of schema, like XML Schema [15], DTD or RelaxNG [11]. XML Schema is a language for XML document format description making it possible to define the structure of a XML document. A XML schema is itself a XML file. Schemas can be employed to guide a XML writer [22] , but they are also employed in different application kinds where the validation is required because the knowledge of the structure of a XML document in particular makes it possible to check the validity of this document. Within the framework of our research, the construction of XML Schema is a remote objective, seeming more difficult to reach at a first glance.

To build the structure of Web documents, this project will initially use the control language of templates named XTiger, and developed by the WAM team at the INRIA [2] French research institute. It presents a new XML language for the template creation making it possible to create documents of any XML target language. While being centered on XHTML documents, the properties of this language allow the creation of valid documents and semantically richer thanks to the use of the micro-formats [19] . This language can be used on the editor of the W3C named Amaya [3], [21].

This language is interesting for our concern, positioning as a light XML schema. It is less constraining, in particular making it possible to structure the document per pieces. In this manner, the parts of the document, for which there is no creation of structure elements, do not block anything in the process of total document structuring. Within this framework, for poorly structured documents, this language will make it possible to overload the document with

components of structures resulting from the statistical elements calculated on the document. We will be easily able to use this language to create document structure components generated automatically starting from the statistical model of the document collection.

4.1 Structure Building

At the level of the document collection, the Data Mining algorithms are applied in order to find a statistical model of the document collection. From this statistical model, it is a question of exploiting the data of the category tree and the elements of probabilities of these categories like their keywords, and their associated probabilities, or their closest or their related categories.

In this manner, we will be able to obtain information on the document structure and be able to produce components like those illustrated on Example 1, 2, or 3.
In the Example 1, the built component is useful to refer the category, called categoryN, to its list of keywords.

Example 1: Component example

```
<xt:component name ="refcategoryN">
  <p class="refcategoryN">
    <xt:repeat minOccurs = nb_keyword_1>
      <span class="keyword_1">
        <xt:use types="string"/>
      </span>
    <xt:repeat minOccurs = nb_keyword_2>
      <span class="keyword_2">
        <xt:use types="string"/>
      </span>
          ...
    <xt:repeat minOccurs = nb_keyword_m>
      <span class="keyword_m">
        <xt:use types="string"/>
      </span>
</xt:component>
```

From this statistical model, we will also be able to obtain information exploiting the data of the category tree. In this manner, we will be able to obtain results on structure using the following XTiger component, Example 2. In this example, a component is created for a *categorie1* and having two sub-categories named *categorie11*, and *categorie12*.

Example 2: Component example

```
<xt:component name ="category1">
  <p class="category1">
    <xt:repeat minOccurs = 1>
      <span class="category11">
        <xt:use types="string">
      </span>
    <xt:repeat minOccurs = 1>
      <span class="category12">
        <xt:use types="string">
      </span>
</xt:component>
```

Unions as *xt:union* component, can be also easily identified while being based on the list of the closest categories. For example, if the categorie number 1 is significantly close to the categories *n*, *2* and *6*, then we can propose the following component (example 3).

Example 3: Union component

```
<xt:union name="category1ref"
          include="refcategoryN  refcategory2  refcategory6">
```

The question of the validity of the created structures will be tackled to make sure that the built structures will be in conformity with the schemas on which they are based. In the case of invalid documents, it will be question of also knowing which solutions to bring. For a XML schema, a validating XML processor could be used.
The behavior of a validating XML processor is highly predictable; it must read every piece of a document and report all well-formedness and validity violations.
In opposition to schema languages: XTiger was not conceived to validate documents. For the XTiger language, the validation is more flexible, and it is a concept which deserves to be thorough within the framework of this work.

4.2 Expected Results

As an example, let us take a Web site of culinary recipies whose web contents are not structured. This kind of Web site is of obvious interest for technologies of the Semantic Web, because the user needs to be able to carry out various research and information extractions.

Example 4: "Ingredients" component example

```
<xt:component name ="Ingredients">
  <p class="Ingredients">
    <xt:repeat minOccurs = 1>
      <span class="Sauce">
        <xt:use types="string"/>
      </span>
    <xt:repeat minOccurs = 0>
      <span class="Miscellaneous">
        <xt:use types="string"/>
      </span>
  </p>
</xt:component>
```

At this step, we will be able to obtain results on the document structure as the illustrated component in Example 4. The element *xt:component* is a constructor which creates a new type containing various elements, being able to contain XTiger component and target language XHTML part. In the example 4, a reference component to a category called *Ingredients* is defined having description elements resulting fromthe lists of its keywords (*Sauce, Miscellaneous*) as done on the previous Example 1.

The Example 5 shows an example of a resulting XML document, which would be structured following the treatment which has just been described. XTiger language helps to describe the structure type of the document. Once the statistical model is completed, we can take these identified elements and put them in the document structure.

Example 5: XML document resulting

```
<Recipe>
 <Usage= "Ingredients">
   <Element> sugar </Element>
   <Element> package cream cheese </Element>
   <SousUsage= "Sauce">
     <Element> salsa cup </Element>
     <Element> salt </Element>
   </SousUsage>
 </Usage>
   ...
 <Usage= "CookingTime">...</Usage>
 <Usage= "Ustensils">...</Usage>
 <Usage= "Directions">...</Usage>
</Recipe>
```

If the nature of the documents to be analyzed is known, as for instance: banks of CVs/Resumes, or electronic message files, and that there exists a corresponding template, we can work starting from its predefined categories (figure 4).
In the example of CVs, these categories could be *Personal information, Education, Professional Experience, and Competences*. The corresponding statistical model is built compared to the document collection to analyze and to these categories. The XTiger template, then the XHTML document are automatically elaborated. In this manner, all the documents are structured in the same way. We can now compare them, query them with XQuery or carrying out a XML transformation.

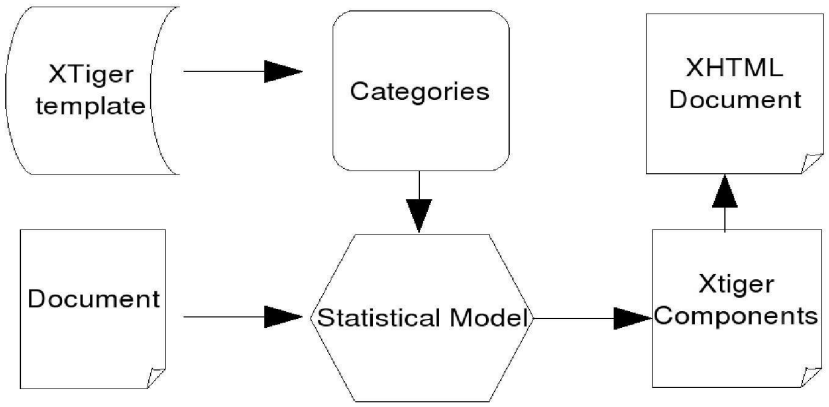

figure 4. Building process starting from an existing template.

The continuity of this work will be done naturally in the automatic building of XML transformations of Web documents while being also based on the data of the statistical models already created for the building of their structures.

5 Applications in bio-data processing

One of the specificities of biology is that the scientific articles are largely available as open files. Moreover, specialized documents, or Web site entireties are not structured and remain not exploitable, as BACTERIO [4] which presents a rich classification of prokaryotes based on the

International Code of Nomenclature of Bacteria (Bacteriological Code). This Web site contains General Considerations, Principles, Rules and Recommendations which govern the way in which the names of prokaryotes are to be used.

These various resources are very heterogeneous in their syntax and their semantics. As a fact, a mine of non structured documents is at our disposal. The process exposed in this research report appears being a good solution to make them exploitable by the machines. These results will be able to apply in the field of molecular biologywhere the explosion of data led to the appearance of many resources accessible on the Web.

The bio-data processing services of the Web are very huge, these various services propose the analysis of the biological data by various programs and Software being based, for example, on research of similarities through large data banks, or on alignments of sequences. These last years, many ontologies appeared in the field of biology. They have a common objective to facilitate the division and the exchange of informations.
Indeed, the predictions and interpretations of data in biology are done by reference and/or comparisons of new data with existing knowledge.

6 Discussion

If we can think that a part of the new information which will be published in the next years on the Web "will be indexed indeed" as of its production in the formalisms of the Semantic Web, (at least, for certain types of more or less factual information and thus easily codable) the major part of information will remain in textual form and then the question of its indexing in the Semantic Web languages will become crucial. It is precisely the objective of the Automatic Treatment of the Languages to understand and to model the human language, to some extent "to make the text calculable". These techniques are useful to reach the contents of the texts, and to make it readable and accessible from the point of view of the machines, our computers.

This work is possible thanks to the emergence of increasingly relevant data analysis methods that give the possibility to build strucrure component such as those in previous sections.
This research report is based on what we already know concerning the Semantic Web. The Semantic Web is an ideal Web and its use would allow a greater control and a broader exploitation of the Web documents. The Web Mining makes it possible to exploit all kinds of data present on the Web in order to include and understand the Web documents uses.

The differences and the gap between the current Web, such as we know it and use it nowadays, and the Semantic Web could be reduced appreciably by applying techniques of data mining for structuring the Web current documents. In fact, the central problems of our Web lie in the lack of formalisms of its contents and this in spite of the constant efforts and the deliveries of XML languages and XML technologies.

We can only note that the efforts of content creation in a valid structure, remain for the user a too great effort. We must thus plan to create an automatic process which makes it possible to structure the Web documents with a minimum of efforts for the users. The process presented in this research report can be applied at the very moment of creation of the Web documents, or on existing
Web documents and deprived of significant structures. This automatic construction of Web documents would make it possible for the current Web to approach its ideal: the Semantic Web. As we saw in this research report, in the general case we do not think of being able to produce a

XML schema. On the other hand, we think that it is reasonably possible to produce a whole range of schemas whose levels of precise details and quality would depend directly on those of the statistical model.

7 Conclusions

Evolution towards Semantic Web appears as inescapable. It is the way of its evolution since the begining of the Web, and also the Computer Science which tender to more and more formalism and structure inside information.
The current Web as we know it today will be deeply modified by this evolution, but it may be done smoothly and slowly.

This research report falls the under work and sets of themes of XML technologies and document adaptation, proposing to build the structure of Web documents which do not have any at the beginning or the structure of which is insufficient. It lies concretely within the scope of XML documents and the problems of the creation of structured documents.

Thanks to techniques of Data Mining, information of structures is captured clarifiing the names of the elements, and certain characteristics of the elements in particular their links, constraints and their logical organization. The automatic process described in this research report will make the current Web more semantic, and will be able to solve applicative problems in particular in bio-data processing.

8 Acknowledgments

We thank Vincent Quint [2] for providing scientific support on this work using XML technologies and particularly the very new XTiger language. We also thank Florence Forbes [5] and Eric Gaussier [6] for giving their technical supports on the Data Mining approach of this research and finally Alain Viari, Anne morgat and Eric Coissac [7] for their discussion on possible applications of our future results in bio-data processing.

9 Bibliography

[1]http://dmoz.org/about.html
[2]http://wam.inrialpes.fr/xtiger
[3]http://wam.inrialpes.fr/software/amaya/.
[4] http://www.bacterio.cict.fr/.
[5] http://mistis.inrialpes.fr/.
[6] http://www-clips.imag.fr/mrim/.
[7] http://www-helix.inrialpes.fr/.
[8] M. Abolhassani, N. Fuhr, and N. Govert. Information extraction and automatic markup for XML documents. pages 159–174. 2003.
[9] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web, may. Scientific American, 45:10, May 2001.
[10] F. Campoy-Flores, V. Quint, and I. Vatton. Microformats and structured editing templates. ACM Symposium on Document Engineering, pages 188–197, October 2006.
[11] J. Clark and M. Murata. RELAX NG specification. ACM Symposium on Document Engineering, December 2001.
[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, B(39):138, 1977.

[13] O. Etzioni. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11):6568, 1996.
[14] S. Eyheramendy, D. Lewis, and D. Madigan. On the naive bayes model for text categorization. the 9th International Workshop on Artificial Intelligence and Statistics, 2003.
[15] D. Fallside and P. Walmsley. Xml schema part 0: Primer, second edition. W3C Recommendation, October 2004.
[16] P. Fankhauser and Y. Xu. MarkItUp! An incremental approach to document structure recognition. In Conference on Electronic Publishing, Document Manipulation and Typography, EP94, Darmstadt, Germany, 1994.
[17] E. Gaussier. Contributions l'accés à l'information documentaire. HDR Université Joseph Fourier, December 2005.
[18]E. Gaussier, C. Goutte, K. Popat, and F. Chen. A hierarchical model for clustering and categorising documents. 24th European Colloquium on Information Retrieval Research (ECIR-02), (2291), 2002.
[19] R. Kosala and H. Blockeel. Microformats: the next (small) thing on the semantic web? IEEE Internet Computing, 10(1):6875, 2006.
[20] L. Ma, J. Shepherd, and A. Nguyen. Document classification via structure synopses. ADC, pages 59–65, 2003.
[21] V. Quint and I. Vatton. Techniques for authoring complex xml documents. ACM Symposium on Document Engineering, page 115123, October 2004.
[22] M. Sifer, Y. Peres, and Y. Maarek. Browsing and editing xml schema documents with an interactive editor. Proceedings of DNIS 2003, 2822:97111, Septembre 2003.
[23] S. Soderland. Learning information extraction rules for semi-structured and free text. Mach. Learn., 34(1-3):233–272, 1999.
[24] G. Stumme, A. Hotho, and B. Berendt. Semantic web mining. state of the art and future directions. Journal of Web Semantics, 4:1–37, 2006.