Indexing a Web Site with a Terminology Oriented Ontology

Emmanuel Desmontils & Christine Jacquin
IRIN, Université de Nantes
2, Rue de la Houssinière, BP 92208
F-44322 Nantes Cedex 3, France
{desmontils,jacquin}@irin.univ-nantes.fr
http://www.sciences.univ-nantes.fr/irin/indexGB.html

Abstract. This article presents a new approach in order to index a Web site. It uses ontologies and natural language techniques for information retrieval on the Internet. The main goal is to build a structured index of the Web site. This structure is given by a terminology oriented ontology of a domain which is chosen a priori according to the content of the Web site. First, the indexing process uses improved natural language techniques to extract well-formed terms taking into account HTML markers. Second, the use of a thesaurus allows us to associate candidate concepts with each term. It makes it possible to reason at a conceptual level. Next, for each candidate concept, its capacity to represent the page is evaluated by determining its level of representativeness of the page. Then, the structured index itself is built. To each concept of the ontology are attached the pages of the Web site in which they are found. Finally, a number of indicators make it possible to evaluate the indexing process of the Web site by the suggested ontology.

keywords : Information Retrieval on the Internet, Web Pages Indexing, Ontologies, Semantic Indexing.

1 Introduction

Searching for information on the Internet means accessing multiple, heterogeneous, distributed and highly evolving information sources. Moreover, provided data are highly changeable: documents of already existing sources may be updated, added or deleted; new information sources may appear or some others may disappear (definitively or not). In addition, the network capacity and quality is a parameter that cannot be entirely neglected. In this context, the question is: how to search for relevant information on the Web more efficiently? Many search engines help us in this difficult task. A lot of them use centralized databases and simple keywords to index and to seek the information. Within such systems, the recall[1] is often relatively high. Conversely, the precision[2] is weak. An intelligent agent supported by the Web site may greatly improve the retrieval process ([4], [1]). In this context, this agent knows its

[1]Recall is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection

[2]Precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved

pages content, is able to perform a knowledge-based indexing process on Web pages and is able to provide more relevant answers to queries. In information retrieval processes, the major problem is to determine the specific content of documents. To highlight a Web site content according to a knowledge, we propose a semi-automatic process, which provides a content based index of a Web site using natural language techniques. In contrast with classical indexing tools, our process is not based on keywords but rather on the concepts they represent.

In this paper, we firstly present the general indexing process (section 2). After having exposed the characteristics of used ontologies (section 3), we will indicate how the representativeness of a concept in a page is evaluated (section 4) and, finally, how this process is evaluated itself (section 5).

2 Overview of the indexing process

The main goal is to build a structured index of Web pages according to an ontology. This ontology provides the index structure. Our indexing process can be divided into four steps (figure 1):

1. For each page, a flat index is built. Each term of this index is associated with its weighted frequency. This coefficient depends on each HTML marker that describes each term occurrence.

2. A thesaurus makes it possible to generate all candidate concepts which can be labeled by a term of the previous index. In our implementation, we use the Wordnet thesaurus ([23]).

3. Each candidate concept of a page is studied to determine its representativeness of this page content. This evaluation is based on its weighted frequency and on the relations with the other concepts. It makes it possible to choose the best sense (concept) of a term in relation to the context. Therefore, the more a concept has strong relationships with other concepts of its page, the more this concept is significant into its page. This contextual relation minimizes the role of the weighted frequency by growing the weight of the strongly linked concepts and by weakening the isolated concepts (even with a strong weighted frequency).

4. Among these candidate concepts, a filter is produced via the ontology and the representativeness of the concepts. Namely, a selected concept is a candidate concept that belongs to the ontology and has an high representativeness of the page content (the representativeness exceeds a threshold of sensitivity). Next, the pages which contain such a selected concept are assigned to this concept into the ontology.

Some measures are evaluated to characterize the indexing process. They determine the adequacy between the Web site and the ontology. These measures take into account the number of pages selected by the ontology, the number of concepts included in the pages... The index is built as a XML file ([28]) and is independent of Web pages.

Our process is semi-automatic. It enables the user to have a global view of the Web site. It also makes it possible to index a Web site without being the owner of these pages. We do not regard it as a completely automatic process. Adjustments should be carried out by the user. The counterpart of this automatisation is, obviously, a worse precision of the process. Lastly, compared to the annotation approach, our indexing process improves information retrieval: it

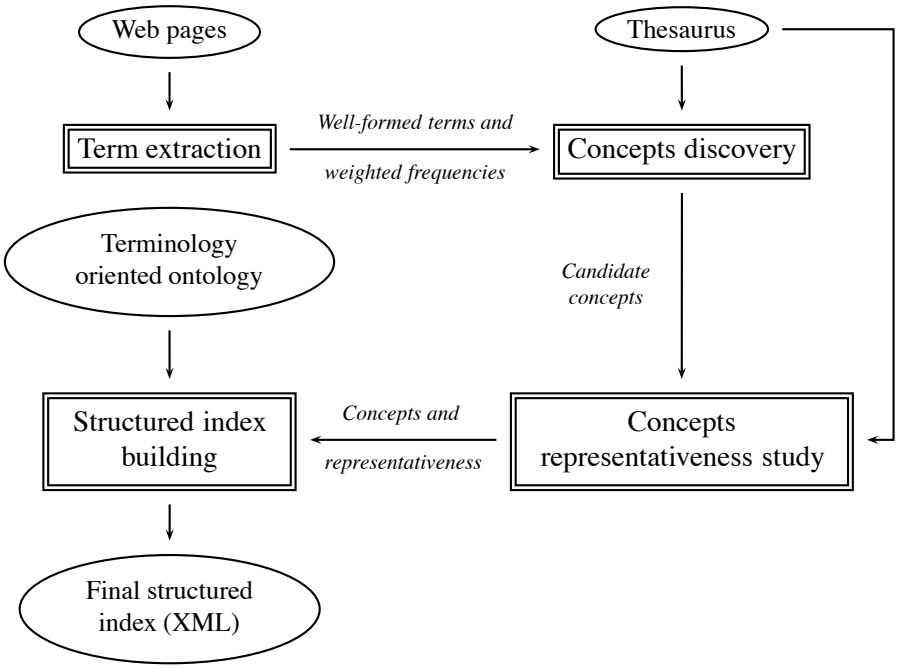

Figure 1: The indexing process

makes it possible to reach directly the pages concerning a concept. By contrast, the annotation approach requires to browse all the pages of the Web site to find this same information. Now, we will study two significant elements: the ontology and the method to evaluate the concepts.

3 Terminology oriented ontologies

3.1 Ontology definition

The term ontology comes from philosophy. In this context, its definition is: «*systematic explanations of the existence*». Furthermore, researchers in Knowledge Engineering give other more suitable definitions with their concerns. In this context, their definitions are strongly dependent on the author's point of view and on his use of ontologies [12, 13]. Some have a formal point of view and work on abstract models of ontologies while others have a more pragmatic approach.

We have chosen this definition of ontology: "*an ontology provides the common vocabulary of a specific domain and defines, more or less formally, terms meaning and some of their relationships*" ([11]). In our context, we thus call ontology a hierarchy of concepts defines in a more or less formal way. For instance, figure 2 shows an extract of the SHOE ontology concerning the American universities.

3.2 Terminology oriented ontology

The concepts of ontologies are usually represented only by a single linguistic term (a label). However, in our context, this term can be at the same time ambiguous (it represents several candidate concepts) and not always unique (existence of synonyms). As a result, within the

```
 <?xml version="1.0" encoding="ISO-8859-1"
      standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="2.1"
          description="...">
  <def-category name="Department"
                isa="EducationOrganization"
                short="university department"/>
  <def-category name="Program"
                isa="EducationOrganization"
                short="program"/>
  <def-category name="ResearchGroup"
                isa="EducationOrganization"
                short="research group"/>
  <def-category name="University"
                isa="EducationOrganization"
                short="university"/>
  <def-category name="Activity"
                isa="SHOEEntity"
                short="activity"/>
  <def-category name="Work"
                isa="Activity"
                short="work"/>
  <def-category name="Course"
                isa="Work"
                short="teaching course"/>
  ...
</ontology>
```

Figure 2: Extract of the SHOE ontology concerning the American universities

framework of texts written in natural language, it is necessary to determine the whole set of the synonyms (candidate labels) to define in a single way a concept. Such process can be found in a manual way in OntoSeek ([14]) or in a semi-automatic way in Mikrokosmos ([25]).

In our context, an ontology is a set of concepts each one represented by a term (a label) and a set of synonyms of this term, and a set of relationships connecting these concepts by the specific/generic relationship, the composition relationship,... Currently, the only relationship we take into account is the "isa" relationship. We call this type of ontology a terminology oriented ontology. Note that our ontologies do not reflect all the inherent aspects to formal ontologies ([11]). Our ontologies are close by their structure to those used in the SHOE project ([21]). Moreover, we choose XML format ([28]) to store our ontologies and our indexing results. The used DTD is rather similar to the SHOE DTD but we made modifications and extensions to this last.

We thus propose a process which makes it possible to determine all the candidate labels of a concept. This process is based on a thesaurus and uses a number of heuristics similar with those proposed by the Mikrokosmos project. The general principle of these heuristics is to try to make a correspondance between the paths according to the "isa" relationship in the

ontology and the paths of hypernyms in the thesaurus. According to the "matching degree", a more or less large confidence is given to such or such set of synonyms (concept). Let us note that experiments using the relationship of composition have not improved the results.

The user can manually finish the disambiguation process of the labels. Indeed, the process can not always select in an unquestionable way the good set of synonyms. The definitions of the sets of candidate synonyms are presented in order to help to this final choice.

However, the process gives results rather satisfactory since it chooses the good sense for nearly 75% of the labels associated with the concepts of the Universities ontology (SHOE project [21]) and for 95% of the label after several modifications (contradictions with the used thesaurus were deleted).

These evaluations were determined with ontologies for which the whole set of the labels associated with the concepts was manually disambiguated. Of course, this disambiguation process depends on the thesaurus used (in our case Wordnet).

4 Index building

The other important part of our process is the indexing process and the evaluation of the importance of a concept in a HTML page. There are two essential steps: (1) terms extraction from Web pages and calculus of the weighted frequency and (2) determination of candidate concepts and the calculus of the representativeness of a concept.

4.1 Terms extraction

The well-formed terms extraction process starts by (1) removing HTML markers from Web pages, (2) dividing the text into independent sentences, and (3) lemmatising words included in the page. Next, Web pages are annotated with part of speech tags using the Brill tagger ([3]). As a result, each word in a page is annotated with its corresponding grammatical category (noun, adjective...). Finally, the surface structure of sentences is analyzed using term patterns (Noun, Noun+Noun, Adjective+Noun...)[7] to provide well-formed terms . For each selected term, we calculate its weighted frequency. The weighted frequency takes into account the frequency of the term and especially the HTML markers which are linked with each of its occurrences. We can notice that the frequency is not a main criterion. Indeed, we work with pages which are of rather restricted size compared to large corpora used in NLP (Natural Language Processing). The influence of the marker depends on its role in the page. For example, the marker "TITLE" will give a considerable importance to the term (*10) whereas the marker "B" (for bold font) has a quite less influence (* 2). The table 1 gives the weight of the most significant markers (the markers weights were determined in an experimental way [10]). In a Web page containing n different terms, for a given term T_i (with i in 1..n), the weighted frequency $F(T_i)$ is determined as the sum of the p weights of HTML markers associated with the p term occurrences. The result is then normalized. This calculus is shown in formula (1) and (2) where $M_{i,j}$ corresponds to the HTML marker weight associated with the jth occurrence of the term T_i.

$$F(T_i) = \frac{P(T_i)}{max_{k=1..n}(P(T_k))} \quad (1)$$

```
<?xml version="1.0" encoding="ISO-8859-1"
      standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="3.0">
  <def-category name="Course" short="teaching course"
                isa="Work">
    <sense name="Course" no="1" origin="WN"
           definition="..." convenience="1.0">
      <synset>class#4,course of instruction#1,
              course of study#2,course#1</synset>
    </sense>
  </def-category>
  <def-category name="Department"
                short="university department"
                isa="EducationOrganization">...
  </def-category>
  <def-category name="University" short="university"
                isa="EducationOrganization">
    <sense name="University" no="3" origin="WN"
           definition="..." convenience="1.0">
      <synset>university#3</synset></sense>
  </def-category>
  <def-category name="Program" short="program"
                isa="Information">
    <sense name="Program" no="4" origin="WN"
           definition="..." convenience="1.0">
      <synset>course of study#1,curriculum#1,program#4,
              syllabus#1</synset></sense>
  </def-category>
  <def-category name="ResearchGroup"
                short="research group"
                isa="EducationOrganization">
    <sense name="ResearchGroup" no="0" origin="TECH"
           definition="" convenience="1.0">
      <synset>research group#0</synset></sense>
  </def-category>
  <def-category name="Activity" short="activity"
                isa="HumanActivity">...
  </def-category>
  <def-category name="Work" short="work"
                isa="Activity">...
  </def-category>...
</ontology>
```

Figure 3: Extract of the terminology oriented ontology concerning the American university

$$P(T_i) = \sum_{i=1}^{p} (M_{i,j}) \tag{2}$$

HTML marker description	HTML marker	Weight
Document title	<TITLE></TITLE>	10
Keyword	<meta name="keywords" ... content=...>	9
Hyper-link	<A HREF=...></A>	8
Font size 7	<FONT SIZE=7></FONT>	5
Font size +4	<FONT SIZE="+4"></FONT>	5
Font size 6	<FONT SIZE=6></FONT>	4
Font size +3	<FONT SIZE="+3"></FONT>	4
Font size +2	<FONT SIZE="+2"></FONT>	3
Font size 5	<FONT SIZE=5></FONT>	3
Heading level 1	<H1></H1>	3
Heading level 2	<H2></H2>	3
Image title	<IMG ... ALT="...">	2
Big marker	<BIG></BIG>	2
Underlined font	<U></U>	2
Italic font	<I></I>	2
Bold font	<B></B>	2
...	...	...

Table 1: Higher coefficients associated with HTML markers

Table 2 shows some results extracted from an experiment on a Web page. Terms are sorted according to the weighted frequency coefficient.

4.2 Page concepts determination

During the term extraction process, well-formed terms and their weighted frequency coefficient were respectively extracted and calculated. The well-formed terms are different forms representing a particular concept (for example “chair”, “professorship”...). In order to determine not only the set of terms included in a page but also the set of concepts in a page, a thesaurus is used. Our experiments use the WordNet thesaurus ([23]). The process to generate candidate concepts is quite simple: from extracted terms, all candidate concepts (all senses) are generated using a thesaurus. A sense is represented by a list of synonym (this list is unique for a given concept). Then for each candidate concept, the representativeness is calculated according to the weighted frequency and the cumulative similarity of the concept with the other concepts in the page. This last one is based on the similarity between two concepts.

We first define the similarity measure between two concepts which makes it possible to evaluate the semantic distance between these two concepts. This measure is defined relatively to a thesaurus and to the hypernyms relationship. In our context, we use the similarity measure defined by [29]. They propose a similarity measure related to the edge distance in the way it takes into account the most specific subsumer of the two concepts, characterizing their commonalities, while normalizing in a way that accounts for their differences. Their measure is shown in formula 3 where c is the most specific subsumer of c_1 and c_2, $depth(c)$ is the number of edges from c to the taxonomy root, and $depth_c(c_i)$ with i in $\{1, 2\}$ is the number of edges from c_i to the taxonomy root through c.

Terms	Weighted frequency
uw	1.00
cse	0.59
uw cse	0.45
computer	0.41
university	0.37
seattle	0.30
article	0.30
science	0.26
research	0.24
professor	0.24
...	...
computer science	0.18
...	...
university of washington	0.16
...	...
program	0.15
...	...
news	0.12
...	...
information	0.09
...	...
message	0.01
...	...

Table 2: Extracted terms and their weighted frequency (sorted according to the weighted frequency). Results coming from http://www.cs.washington.edu/news/

$$sim(c_1, c_2) = \frac{2 * depth(c)}{depth_c(c_1) + depth_c(c_2)} \tag{3}$$

This measure performs a little worse than the Resnik's measure ([26]) but better than the traditional edge-counting measure (see related works for more details).

For evaluating the relative importance of a concept in a page, we define its cumulative similarity. The cumulative similarity measure associated with a concept in a page, noted $\widehat{sim}$, is the sum of all the similarity measures calculated between this concept and all the other concepts included in the studied page. In this formula, a specific concept is unified with the corresponding synset (set of synonyms) in WordNet. The measure is shown in formula 4[3], where l_k synsets are associated with a term T_k, and there are m terms in the studied Web pages.

$$\widehat{sim}(synset_i(T_k)) = \sum_{j \in [1,k-1] \cup [k+1,m]} \sum_{l=1}^{l_j} sim(synset_i(T_k), synset_l(T_j)) \tag{4}$$

In this calculus, all similarities are not been taken into account in order to discriminate the results: a threshold is applied. Finally, we determine a representativeness coefficient which

[3] $\widehat{sim}(synset_i(T_k))$ is normalized

determines the representativeness of a concept in a document. The coefficient is a linear combination of the weighted frequency and of the cumulative similarity of a concept (formula 5)[4]. This coefficient is the major one to qualify answer to a request. The empirical values for α and β are respectively 2 et 1.

$$representativeness(synset_i(T_k)) = \frac{\alpha * F(synset_i(T_k)) + \beta * \widehat{sim}(synset_i(T_k))}{\alpha + \beta} \quad (5)$$

The table 3 shows the effect of the representativeness on the concepts order (terms found in the page are in bold font). Some concepts are higher in the table 3 than in the table 2. For instance, news#1 (weighted frequency 0.12, representativeness 0.51) or information#1 (weighted frequency 0.1, representativeness 0.59). This is a good result for a page related to a news page. If we analyse the result more in details, the concepts: news#4 and news#2 have a representativeness equal to 0.49. This is not very different from the degree of news#1 which is equal to 0.51. The explanation is that Wordnet includes too much fine-grained sense distinctions. In fact, in the thesaurus, the three previous concepts have all the same subsumer. Then, an automatic process cannot distinguish these three concepts. Wordnet was built by linguist and is not always effective in NLP [25].

5 Associating concepts and synsets

At this point, we have on the one hand a terminology oriented ontology and on the other hand candidate concepts with their representativeness coming from HTML pages. In the next step, candidate concepts are matched with concepts of the ontology. If a concept is in the ontology and in a Web page, the URL of this page and its representativeness are added to the ontology.

To evaluate the appropriateness of an ontology according to a set of HTML pages, five typical coefficients are calculated. These coefficients are normalized. The first four coefficients define:

- the rate of concepts directly involved in HTML pages, called *the Direct Indexing Degree or DID*;
- the rate of concepts indirectly involved in HTML pages (calculated by the way of the generic/specific relationship), called *the Indirect Indexing Degree or IID*;
- the rate of pages concerned with the ontology concepts, called *the Ontology Cover Degree or OCD*, which gives the number of Web pages that involve at least one concept of the ontology;
- the Mean of the Representativeness of the candidate Concepts (MRC).

These coefficients (DID, IID, OCD, MRC) are evaluated for different thresholds applied on the representativeness (0 to 1 with a step equals to 0.02). For each coefficient its weighted mean (WM) is calculated. For instance, formula 6 presents the calculus of the weighted mean for the direct indexing degree (DID).

[4]The representativeness is normalized. $F(synset_i(T_k)$ is the normalized sum of all the weighted frequency related to $synset_i(T_k)$.

Concepts	Weighted frequency	Representativeness
uw#0	1.0	1.0
award#2, accolade#1, honor#1, honour#2, laurels#1	0.20	0.7
computer#1, data processor#1, electronic computer#1, information processing system#1	0.41	0.68
information#1, **info**#1	0.1	0.59
cse#0	0.59	0.59
university#2	0.37	0.58
course of study#1, **program**#4, curriculum#1, syllabus#1	0.15	0.53
calculator#1, reckoner#1, figurer#1, estimator#1, **computer**#2	0.41	0.51
news#1, **intelligence**#4, tidings#1, word#3	0.12	0.51
news#2	0.09	0.49
news#4	0.09	0.49
voice#6	0.01	0.51
voice#2, vocalization#1	0.01	0.51
message#2, **content**#2, subject matter#1, substance#6	0.01	0.51
language#1, linguistic communication#1	0.01	0.51
article#3, clause#2	0.30	0.5
submission#1, **entry**#4	0.01	0.5
subject#1, topic#1, theme#1	0.01	0.5
university#3	0.37	0.42
...	...	...

Table 3: Extracted concepts after the calculus of the representativeness degree (sorted according to the representativeness). Results coming from http://www.cs.washington.edu/news/

$$\overline{DID_{s,o}} = \sum_{i=0}^{1} (i * DID_{i,s,o}) \tag{6}$$

This calculus privileges the concepts which are more representative of the pages. A representative ontology of a site has the weighted mean nearly equal to 1. This evaluation depends on the thesaurus used because it depends on the used relationships. Finally, the global evaluation of the indexing process (OSAD: Ontology-Site Adequacy Degree) is a linear combination of these weighted means. Currently, the coefficients are evaluated in an experimental way. The equation 7 gives the present evaluation where s is a Web site and o an ontology. The experiment shows that a value of 0.3 for the representativeness gives good results. Below this threshold, too many concepts with a low representativeness are kept. For this threshold, the discrimination of concepts is relatively effective (the larger the Web pages are, the more effective the process is).

$$OSAD_{s,o} = \frac{\overline{IID}_{s,o}}{2} + \overline{DID}_{s,o} + 2 * \overline{OCD}_{s,o} + 2 * \overline{MRC}_{s,o} \tag{7}$$

The figure 4 presents indexing results related to the Web site: "http://www.cs.washington.edu/" (1315 HTML pages). This is the site of the department of computer science of the Washington university. It was chosen because of its a priori adequacy with our ontology. However, the

Ontology-Site Adequacy Degree (OSAD) is not very high (56%). The explanation is that the used ontology (the SHOE ontology with some extensions and modifications) does not cover all the studied domain. For instance, the studied site has numerous personal Web pages which are rarely indexed by the ontology. Figure 5 presents an extract of the structured index.

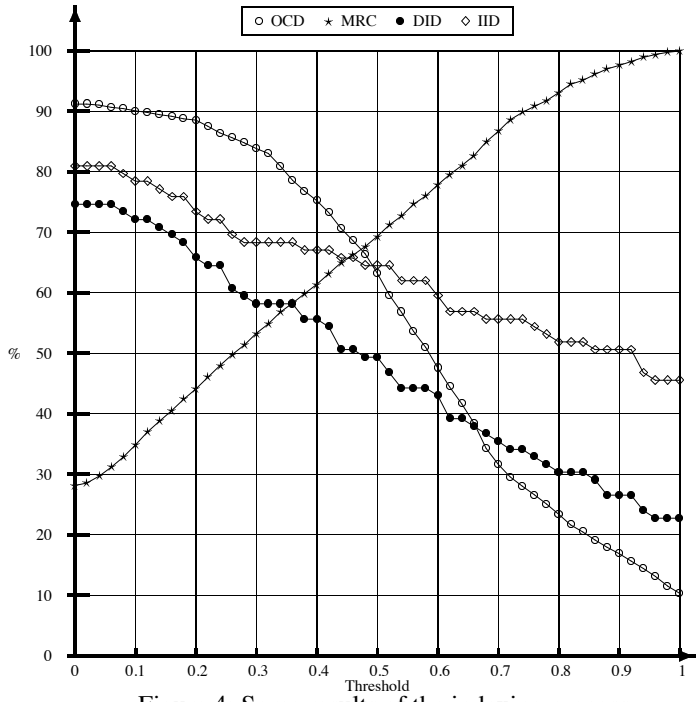

Figure 4: Some results of the indexing process

The indexing process can highlight concepts, which do not match with concepts of ontologies. In this case, we may search for ontologies related to this index. In the future, we will be able to start again the indexing process when the content of the site evolves or when ontologies are updated. This process can only be executed on modified pages.

The evaluation process enables us to evaluate the adequacy between the pages of the site and the ontology and thus to adopt various strategies depending on the coefficients value:

1. the coefficients are correct: the structured index is kept and exploited;
2. the coefficients are not correct:
   (a) the pages which are not suitable are deleted (the OCD and/or the MRC coefficient are low);
   (b) the ontology is updated (the DID coefficient is low);
   (c) a new ontology is chosen and the index is built again (the whole set of coefficients is low);

6 Exploitation of our approach for query answering

Most of search engines use simple keywords to index web pages. Queries are often made up of a list of keywords connected by logical operator (“and”, “or”...). In our context, we

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="3.0" description="">
 <def-category name="University" short="university"
              isa="EducationOrganization">
   <sense name="University" no="3" origin="wn" convenience="1.0">
     <synset>university#3</synset>
     <page name="http://www.cs.washington.edu/info/contact/"
           frequence="0.5" representativeness="0.4"/>
     <page name="http://www.cs.washington.edu/info/aboutus/"
           frequence="0.54" representativeness="0.49"/>
     <page name="http://www.cs.washington.edu/education/courses/590m/"
           frequence="0.4" representativeness="0.4"/>
     <page name="http://www.cs.washington.edu/outreach/"
           frequence="0.28" representativeness="0.34"/>
     <page name="http://www.cs.washington.edu/mssi/"
           frequence="0.5" representativeness="0.43"/>
     <page name="http://www.cs.washington.edu/general/overview.html"
           frequence="0.87" representativeness="0.81"/>
     <page name="http://www.cs.washington.edu/education/courses/599/"
           frequence="0.4" representativeness="0.35"/>
     <page name="http://www.cs.washington.edu/workforce/tnt/"
           frequence="0.25" representativeness="0.35"/>...
    </sense>
 </def-category>
 <def-category name="Department" short="university department"
              isa="EducationOrganization">
   <sense name="Department" no="1" origin="wn" convenience="1.0">
     <synset>department#1,section#11</synset>
     <page name="http://www.cs.washington.edu/education/courses/444/"
           frequence="0.29" representativeness="0.31"/>
     <page name="http://www.cs.washington.edu/lab/facilities/la2.html"
           frequence="0.5" representativeness="0.41"/>
     <page name="http://www.cs.washington.edu/ARL/"
           frequence="0.21" representativeness="0.32"/>
     <page name="http://www.cs.washington.edu/"
           frequence="0.33" representativeness="0.37"/>
     <page name="http://www.cs.washington.edu/desktop_refs.html"
           frequence="0.5" representativeness="0.43"/>
     <page name="http://www.cs.washington.edu/news/jobs.html"
           frequence="0.44" representativeness="0.46"/>
     <page name="http://www.cs.washington.edu/admin/newhires/faq.html"
           frequence="0.29" representativeness="0.32"/>
     <page name="http://www.cs.washington.edu/info/videos/index.html"
           frequence="0.39" representativeness="0.38"/>
     <page name="http://www.cs.washington.edu/affiliates/corporate/"
           frequence="0.5" representativeness="0.51"/>...
   </sense>
 </def-category>...
</ontology>
```

Figure 5: Extract of the structured index based on the terminology oriented ontology concerning the American universities

use the terminology oriented ontology and the structured index in order to improve the query answering process. Queries are not only processed at the terminological level but also at the conceptual level. This approach provides several improvements:

1. a user's query is expanded: terms are transformed into concepts;
2. logical operators have a richer semantics than in the simple keywords world;
3. the answers are more suitable to a user's query.

The query expansion is thus improved by the use of ontologies. Often, when a user proposes a query which contains terms connected by logical operators, these terms are often ambiguous. In our approach, terms are replaced by their associated concepts. The candidate concepts are first selected in the ontology. Then, the other concepts of the query and the logical operators are studied. Finally, if a term is still associated with several candidate concepts, the user's assistance is required. If set of query terms are not associated with any concepts at the end of this process, they are regarded as not relevant for the site. According to the logical operator, either they are suppressed from the query or the query has no response.

The conceptually expanded query can be exploited to seek pages corresponding precisely to its content. The ontology makes it possible to improve the interpretation of the used logical operators. Currently, in one hand, the "and" and "or" operators have the same interpretation as in the traditional keywords approach. In the other hand, the "no" and "near" operators have a different semantics. For a query containing a "no" operator, we add to the concerned concept, all the concepts which are more specific than this concept according to the "isa" relationship. So, all pages containing these concepts will be rejected. The "near" operator is not related to the distance between words (number of words between two words) as in the classical approach. But, it is related to a semantic distance between concepts according to the similarity measure [29] used to calculate the representativeness coefficient. In our context, the "near" operator becomes an unary operator and makes it possible to add to the query all the concepts semantically connected to the targeted concept and in its neighborhood.

7 Related works

Our choices differ from related works especially from work on annotation of Web page like KA2 ([9], [2]), SHOE ([21]) or WebKB([22]). These two projects annotate manually Web pages using semantic tags. SHOE proposes a set of *Simple HTML Ontology Extension* to annotate Web pages with ontology-based knowledge concerning page contents. In this context, an agent can use this knowledge to manage effectively information requests.

In all the cases, the goal is to use semantic information to improve the information retrieval. However, in these approaches, annotations are strongly linked to document. The author of pages progressively indicates handled knowledge where it appears. The problem is that any modification or new generation of the pages requires to remake entirely or partly the annotations. Nevertheless, the precision of this process is extremely fine. Moreover, the methods based on annotation are manual or semi-manual (an user interface helps the user to annotate the document [16]). Therefore, they are very time expensive and can be carried out only by specialists ([15]).

However, this manual process is time expensive, complex, and information and knowledge are mixed. The information management difficulty is thus increased ([15]). In addition,

semantically annotated documents are not today and perhaps may be never available on the Web. These two projects work on restricted domain and scaling up to the entire Web is a titanic task ([15]). Moreover, in this context, all Web page builders have to accept to annotate their own pages. The consensus needed by this protocol is far to be widely admitted and is at the opposite of the Web philosophy. Another project is the "WebKB" project ([22]). It proposes another manual process to annotate Web pages using an ontology represented with a conceptual graph ([27]), which is built using a linguistic thesaurus. Even if the used language is different from the two previous projects, annotations are also included in the HTML pages. Moreover, the thesaurus is only used to extend the ontology. It is not used to automatically index natural language documents.

Like in OntoSeek project ([14]), our approach adds linguistic attributes to ontologies using the WordNet thesaurus to improve our semi-automatic Web site knowledge discovery. Guarino calls this process a *disambiguation process*. However, the manual process OntoSeek uses ontologies not to define the knowledge of a Web site but to find user's data in a large classical database of Web pages. Another project proposes a similar process: the Mikrokosmos project ([25]) to provide a knowledge base for machine translation process. This process is another semi-automatic process (the user can improve manually the disambiguation results). It studies several heuristics. The most important are an hierarchical heuristics and a similarity heuristics. The hierarchical heuristics uses the generic/specific relationship in the ontology and the hypernyms relationship in the thesaurus. For [25], the hierarchical heuristics seems to be the more effective to select senses. Therefore, we choose to use this heuristics and to improve it.

Some projects of the KDD (Knowledge Discovery in Databases) community are interested by extracting knowledge from Web sites. [8] apply techniques of KDD to keywords which are attached to the documents and which are then regarded as attributes. These mining techniques use statistical analysis to discover association rules and interesting patterns over keywords distributions and associations. Other researchers [18] use terms automatically extracted from documents to characterize the document and to find associations which connect the terms to the documents. Another approach is to apply KDD techniques after the use of information extraction techniques, which transform information located in texts into a structured database [6]. Other approaches [20] mixe NLP techniques and KDD techniques to extract automatically information from documents. They do not use keywords as attribute but use concepts which are acquired by the way of a thesaurus. The approach of the last authors seems the most interesting because they do not work any more with simple keywords but with the concepts included in documents. Compared to KDD techniques like [20], we also work on conceptual level instead on the simple keywords level. But we take the option to have linguistic processing much finer and especially we privilege an a priori knowledge on the studied domain (one or several ontologies). [20] use a priori knowledge on the studied domain (a thesaurus) exclusively to extract the concepts of the pages. In our approach, the concepts are also extracted from the pages using a thesaurus, but the indexing process itself is also based on an ontology of the domain. [24] asserts besides that for an effective extraction of knowledge, a priori knowledge on the studied domain (for example ontologies) is essential.

Many measures of similarity are defined in related works. For [19], the information shared by two concepts is indicated in an "isa" taxonomy by the most specific concept that subsumes them. The semantic similarity of two concepts in a taxonomy is the distance between the nodes corresponding to the items which are compared (edge-counting). The shorter the path from one node to another is, the more similar they are. Given multiple paths, one takes the

length of the shortest one.

A widely acknowledged problem ([26]) with this approach is that it relies on the notion that links in the taxonomy represent uniform distances (but it is most of the time false). [26] describes an alternative way to evaluate semantic similarity in a taxonomy based on the notion of information content. All links in a taxonomy are weighted with an estimated probability (concept occurrences in corpora), which measures the information content of a concept. The main idea is: the more concepts share information, more similar they are. The information shared by two concepts is indicated by the information content of the concepts that subsumes them in the taxonomy. The probability P of a concept c is based on the probability associated with the concept plus the probability associated with all its descendant concepts. $P(c)$ is then used to calculate the information content of a concept c which is equal to $-log(P(c))$.

8 Conclusions

In this paper, we have presented a semi-automatic process to index a Web site by its content. This process builds a structured index coming from an ontology and pages of a Web site. After the construction of a flat index where all terms have a weighted frequency, we determine candidate concepts associated with these terms. For each concept, a representativeness coefficient is calculated. Finally, the most representative concepts in a Web page are selected, and those which belong to the ontology are kept. The final structured index is organized according to the ontology. With each ontology concepts a set of Web pages is associated from where the potential concepts were extracted.

This process comprises a number of advantages on the traditional indexing methods (only based on keyword retrieval) and even on the methods of Web site annotation:

1. selected pages contain not only the keywords but also the required concepts ;
2. these concepts are representative of the topics treated in selected pages ;
3. terms which are responsible of the page selection are not always those of the request but can be synonyms ;
4. pages can comprise not only the required concepts but also more specific ones ;
5. the importance of a concept depends not only on its term frequency but also on the HTML markers which describe it and on its relations with the other concepts of the page...

The indexing process can be used not only for retrieving information but also for valuing the appropriateness of a Web site with regard to a domain or a knowledge. This latter case enables us to classify a Web site in a hierarchical index of a classical search engine (Yahoo !, Excite...). Note that such hierarchies can be themselves considered as general ontologies ([17]).

Currently, other Web sites on American universities are indexed in order to compare their results to those of the Washington university. In order to improve the indexing results, we may also improve the coverage degree of the ontology on our studied domain. We study also other relationships than the generic/specific relationship in order to improve the process of concepts extraction. We have developed a measure according to the composition relationship, but we must also evaluate it in an experimental way.

The results presented in this paper can be used in various applications. They are currently being incorporated within the Bonom Multi-agent system ([5], [4]) to search for relevant information on the Internet. The system involves different types of agents among which "site agents" which encapsulate information sources. The methods we propose are implemented within the site agents. They greatly improve the site analysis process and the query answering process.

References

[1] N. Ashish and C. A. Knowblock, "Semi-Automatic Generation Internet Information Sources", In 2nd IFCIS Conference on Cooperative Information Systems (CoopIS), Charleston, SC, 1997.

[2] V. R. Benjamins, D. Fensel, A. Gomez-Perez, S. Decker, M. Erdmann, E. Motta, and M. Musen. "Knowledge Annotation Initiative of the Knowledge Acquisition Community KA2". In Proceedings of the 11th Banff knowledge acquisition for knowledge-based system workshop, Banff, Canada, 1998, pp. 18-23.

[3] E. Brill, "Transformation-based error-driven learning and natural language processing: a case study in Part-of-speech Tagging". Computational Linguistics, vol. 21, 1995, pp. 543-565.

[4] S Cazalens, E Desmontils, C Jacquin, and P Lamarre, "A Web Site Indexing Process for an Internet Information Retrieval Agent System", International Conference on Web Information Systems Engineering (WISE'2000), IEEE Computer Society Press, Hong-Kong, 19-20 June, 2000, pp. 245-249.

[5] S. Cazalens and P. Lamarre, "An organization of Internet agents based on a hierarchy of information domains", In Proceedings MAAMAW, Yves Demazeau and Francisco J. Garijo editors, may 2001

[6] J. Cowie and W. Lehnert. "Information extraction". In Communications of the ACM, number 1, volume 39, january 1996.

[7] B. Daille, "Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques", PHD Thesis, Paris 7, 1994.

[8] R. Feldman and I. Dagan. "Knowledge discovery in textual databases (KDT)". In First international conference on knowledge discovery (KDD'95), Montreal, august 1995.

[9] D. Fensel, S. Decker, M. Erdmann, and R. Studer. "Ontobroker: Or How to Enable Intelligent Access to the WWW". In Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based System Workshop (KAW'98), Banff, Canada, 1998.

[10] J. Gamet. "indexation de pages web". Rapport de DEA informatique, université de Nantes, 1998.

[11] A. Gomez-Perez. "Développements récents en matière de conception, de maintenance et d'utilisation des ontologies". In Proceedings of colloque Terminologie et intelligence artificielle de Nantes, 10-11 mai 1999, revue terminologies nouvelles, pp. 9-20.

[12] T. Gruber, "A Translation Approach to Portable Ontology Specification". In Knowledge Acquisition journal, vol 5, pp 199-220, 1993.

[13] N. Guarino. "Some Organizing Principles for a Unified Top-Level Ontology". In Spring Symposium series on ontological engineering, pp 57-63, 1997

[14] N. Guarino, C. Masolo, and G. Vetere, "OntoSeek: Content-Based Access to the Web", IEEE Intelligent Systems and Their Applications, Elsevier Science, 14(3), 1999, pp. 70-80.

[15] J. Heflin, J. Hendler, and S. Luke, "Applying Ontology to the Web: A Case Study", In International Work-Conference on Artificial and Natural Neural Networks (IWANN), 1999.

[16] J. Kahan, M. Koivunen, E. Prud'Hommeaux and R.R. Swick "Annotea: An Open RDF Infrastructure for Shared Web Annotations", In proceedings of the WWW'10 conferences, Hong Kong 2001

[17] Y. Labrou and T. Finin, "Yahoo! as an Ontology - Using Yahoo! Categories to Describe Documents", In Proceedings of CIKM'99, Kansas City, MO, Oct. 1999, pp. 180-187.

[18] S. Lin and al. “Extracting classification knowledge of internet documents with mining term associations: a semantic approach”. In International ACM-SIGIR conference on research and development in information retrieval (SIGIR-98).

[19] J. H. Lee, M. H. Kim, and Y. J. Lee, “information retrieval based on conceptual distance in IS-A hierarchies”, journal of documentation, 49(2) , 1993, pp 188-207.

[20] S. Loh, L.k. Wives and J Palazzo M de Oliveira. “Concept-based knowledge discovery in texts extracted from the web”. In journal SIGKDD explorations, number 1, volume 2, pp 29-39, 2000.

[21] S. Luke, L. Spector, and D. Rager. “Ontology-Based Knowledge Discovery on the World-Wide-Web”. In Proceedings of the workshop on Internet-based information system, AAAI’96, Portland, Oregon, 1996.

[22] P. Martin and P. Eklund, “Embedding Knowledge in Web Documents”, In Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, May 11-14, 1999 (http://www8.org).

[23] G. A. Miller, “WordNet: an Online Lexical Database”, International Journal of Lexicography, 3(4), 1990, pp. 235-312.

[24] D. Mattox, L. Seligman and K. Smith, “Rapper: a wrapper generator with linguistic knowledge”, In ACM workshop on information and data management, 2000.

[25] T. O’Hara, K. Mahesh, and S. Niremburg, “Lexical Acquisition with WordNet and Microkosmos Ontology”, workshop on “usage of WordNet in natural language processing systems”, 8 pages, Coling-ACL’98

[26] P. Resnik, “Semantic similarity in a taxonomy : an information-based measure and its application to problems of ambiguity in natural language”, journal of artificial intelligence research, 11, July 1999, pp. 95-130.

[27] J. F. Sowa, “Conceptual Structures, Information Processing in Mind and Machine”, Addison Wesley Publishing Company, 1984

[28] W3C. “Extensible Markup Language (XML) 1.0”. W3C Recommendation, Reference: REC-xml-19980210, 10 February 1998, http://www.w3.org/TR/REC-XML

[29] Z. Wu and M. Palmer, “verb semantics and lexical selection”, In Proceedings of the 32nd annual meeting of the association for computational linguistics, Las Cruses, New Mexico, 1994

[30] B. Yuwono and D. L. Lee. “WISE: A World Wide Web Resource Database System”. IEEE Transactions on Knowledge and Data Engineering, 8(4), 1996, pp. 548-554.

To appear in :
“The Emerging Semantic Web”,
I.F. Cruz et al. (Eds),
IOS Press,
pp.181–197,
2002