Web Site Indexation and Ontologies

E. Desmontils & C. Jacquin

Institut de Recherche en Informatique de Nantes
2, rue de la Houssinière
B.P. 92208
F-44322 NANTES CEDEX 3

— Information Retrieval and Ontologies —

RESEARCH REPORT

N° 00.12

November 2000

IRIN, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3

E. Desmontils & C. Jacquin

Web Site Indexation and Ontologies

28 p.

Les rapports de recherche de l'Institut de Recherche en Informatique de Nantes sont disponibles aux formats PostScript® et PDF® à l'URL :

http://www.sciences.univ-nantes.fr/irin/Vie/RR/

Research reports from the Institut de Recherche en Informatique de Nantes are available in PostScript® and PDF® formats at the URL:

http://www.sciences.univ-nantes.fr/irin/Vie/RR/indexGB.html

Web Site Indexation and Ontologies

E. Desmontils & C. Jacquin

Abstract

This report presents a new approach to index a web site using ontologies and natural language techniques for Internet information retrieval. Ontologies are used to index a web site and, as a result to represent the web pages content. First, a linguistic ontology (a thesaurus) is used to disambiguate the label of ontology concepts. This disambiguation process uses several hierarchical heuristics that take advantage of the "isa" relationship on both the ontology and the thesaurus. Natural language techniques based on extraction of well-formed terms are also presented. They used a web page particularity: HTML markers. Then, from the previous extracted terms, associated concepts are generated using a linguistic ontology and a semantic similarity measure. At each candidate concept is associated a couple of coefficient: the convenience coefficient and the weighted frequency coefficient. The match between web page concepts and ontology concepts is then presented. Moreover, results about different web sites on different domains are presented. They point out the good accuracy of our ontology indexation process.

Categories and Subject Descriptors: H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Indexing methods, Linguistic processing, Thesauruses*; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Search process*

Additional Key Words and Phrases: World Wide Web, Information Retrieval, Internet, Indexing Web Pages, Ontologies, Semantic Indexation, Disambiguation Process

Contents

List of Figures

List of Tables

1 Introduction

Searching for information on the Internet means accessing multiple, heterogeneous and distributed 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 entirely neglected. In this context, the question is: how to search relevant information on the web more efficiently? Many search engines help us in this difficult task. A lot of them use centralized database and simple keywords to index and to access the information. With such systems, the recall is often rather convenient. Conversely, the precision is weak. Some works in the multi-agent community (Yuwono and Lee, 1996; Ashish and Knoblock, 1997; Cazalens *et al*., 2000) show that an intelligent agent supported by the web site may greatly improve the retrieval process. In this context, this agent knows its pages content and is able to answer or not to queries. Namely, web sites become "intelligent" and are able to perform a knowledge-based indexation process on web pages. The work presented in this paper is related to this framework i.e. how, a web site can know its web page content and how ontologies can be used for information retrieval purpose? Our project, called *Thoth*, plans to respond as far as possible to these questions.

Keywords, which are used to index a site in usual approach, are natural language terms, which are basically ambiguous. Working at the word level is not enough to disambiguate a query (Gonzalo *et al*., 1998) or to index a document (Luke *et al*., 1996; Fensel *et al*., 1998). To improve the information retrieval process on the web, we may work at the conceptual level. To this end, ontologies are often used. They are concept hierarchies, which provide the common vocabulary of a specific domain. For instance, a hierarchy as *Yahoo!* could be considered as a simple ontology (Labrou *et al*., 1999). In information retrieval processes, these ontologies are used in natural language context (web page are written in natural language). For this reason, an ontology concept is represented by a label which is often a term. These labels are the bridge between words in a page and associated ontology concepts. Therefore, at a label may correspond a single meaning. However, in natural language a term could have several meanings and a meaning could be represented by several terms. Therefore, a linguistic ontology can help to this disambiguation problem. In addition, extracting precise indexes from pages, which could represent the web page content, can improve the retrieval process. Indeed, web pages are written in natural language and included concepts can appear on different forms (synonyms for example). Natural language techniques, as terminological extraction and measuring of word similarities, can greatly help to collect all forms and to determine the associated concepts.

In this research report, we present the main process characterizing a web site content (section 2). Next, to detail this process, we present:

- The features of Thoth's ontologies, and the process to disambiguate concept label (section 3). This disambiguation process uses a linguistic ontology and some hierarchical heuristics. Experimental results are then presented and commented.

- The natural language process to extract well-formed terms from pages, and the generation of concepts associated with a page (section 4). Each of them is weighted by a convenience coefficient (which is determined by semantic similarity techniques) and weighted frequency coefficient (which points out the weight of this concept in the page).

- The match between the ontology concept and the extracted concepts from pages (section 5). At least, before concluding we present some experimental results proceeded on three different sites and we discuss them.

2 Web site content highlight

In information retrieval processes, the major problem is to determine the specific content of documents. To highlight a web site content according to an ontology, we propose a semi-automated process, which provides a partial indexation by content of a web site using natural language techniques. We argue that such a process is necessarily semi-automatic. The user is the only way to finalize the process to correct errors and sometimes to complete it with manual extensions concerning specialized concepts.

In related works, distributed ontological and linguistic-based web site indexation is rather little used. WISE (Yuwono and Lee, 1996) proposes an *indexer robot* to build a centralized index database. This index process is based on statistical methods (Ranking algorithm TFxIDP) without using natural language term relations and the site knowledge. (Ashish and Knoblock, 1997) describe an *information mediator*. This agent uses *wrappers* to index web sites only using two HTML tags (font size and indentation spaces). Nevertheless, this solution provides consistent results only for stable structured information sources. The "Web->KB" system (Craven *et al*., 2000) uses a knowledge database built by a training process (a supervised learning process) based on an ontology and web pages examples. However, first it does not use natural language techniques (especially, like most of works, it does not consider synonyms of concepts, complex text term relations...) and second, its training step causes to rebuild the databases each time the site adds new pages.

Nearest ontology-based works relative to the web site knowledge definition are the "SHOE project" (Luke *et al*., 1996) and the Ontobroker's framework web page annotation (KA2 project; Fensel *et al*., 1998) to manually annotate web pages using semantic tags. The first one is less compact and more understandable than the second one. Conversely, SHOE is less powerful concerning concepts relation management than KA2. SHOE proposes a set of *Simple HTML Ontology Extensions* to annotate web pages with ontology-based knowledge concerning page contents. Then, an agent can use this knowledge to manage powerful information requests. Since it is *a manual process*, annotations are accurate and relevant. However, this manual process is time expensive, complex, and information and knowledge are mixed up. The information management difficulty is thus increased (Heflin *et al*., 1999). In addition, semantically annotated document 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 may be a titanic task (Heflin *et al*., 1999). Moreover, in this context, all web page builders have to accept to annotate these 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 (Martin and Eklund 1999). It proposes another manual process to annotate web pages using a wide ontology represented with a conceptual graph (Sowa, 1984), which is built using a linguistic thesaurus (Martin, 1995). 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 not automatically to index and to search in natural language data.

Inputs of this ontology based indexation process are (1) typical ontologies concerning the knowledge to highlight and (2) an HTML page set of a Web site. Briefly, an ontology is a set of concepts which are connected with relations. Our ontologies are enriched using a thesaurus to associate to each concept label the set of its possible synonyms. The section 3 will give a detail account of what is an ontology and how we can obtain it. Ontologies are used to index the site (figure 1), that is to say we plan to associate, with each concept of the ontology, pages were they could be founded. We can call this process the ontology indexation of web pages. Namely, we first extract indexes that point out the site content from the HTML pages using natural language based techniques. An index of a web page is a term (a set of words), which can be a significant concept of this page. Second, we try to define concepts they could represent (these two last steps are described in section 4). Then, we match them with concepts of selected ontologies and a set of measures is computed to evaluate this process (section 5). Finally, either the process is ended (measures are convenient for the user) or another indexation process is started with a new ontology (if indexation results does not suit him/her).

Note this process can be applied not only to web pages but also to news and mailing list databases. The scope of this paper is limited to HTML pages.

Our process does not change web site data (html pages, news archives...) but create an additional XML file (W3C, 1998). This file refers to a DTD close to the one used in the SHOE project (Luke *et al*., 1996). The result of our process is composed of ontologies that refer to web pages. These ontologies are also described in a separated XML file. Then, HTML pages contain only owner's information. Therefore, they are easy to manage without considering the added knowledge. In addition, we can rebuild the added knowledge without modifying web pages. Finally, every web browsers can display HTML pages without the annotation that risks to provide problems. Therefore, our process is not an annotation process (concepts of the ontology are inserted in the text) but an indexation process (ontology concepts point out web pages where they are). This allows us to access directly to concerned pages searching a given concept and not to parse all web pages of the site. It also allows us to characterize answers using a coefficient of convenience.

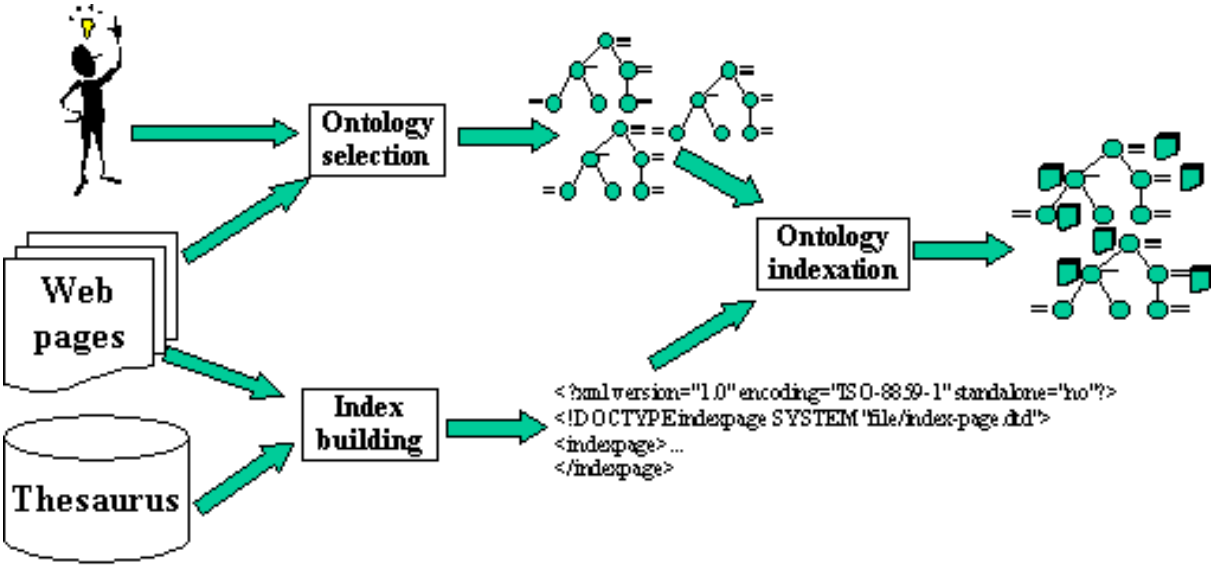

Figure 1: The general process

3 Ontologies

3.1 Definition

Many definitions of ontology are available in related works. In our context, we have chosen this one: "*an ontology provides the common vocabulary of a specific domain and defines, more or less formally, terms meaning and some of their relationships*" (Gomez-Perez, 1999). From a Knowledge Engineering point of view, an ontology is a set of concepts represented by a label, and a set of relations connecting these concepts. Major relationships are the "isa" relationship, the "partOf" relationship...

We have chosen XML to store our results, and our ontologies. The SHOE DTD seems convenient for us but it lacks some important information we will present later. So, we have extended this DTD to add our information. Our ontology DTD is described in the appendix 1. Currently, we have only managed the "isa" relationship but other ones can be used (and have to be used). For example, an extract of the Thoth's University ontology is shown in the appendix 2.

A major problem that concerns ontologies is: where to find them and how to build them? First, there exist ontologies for specialized field of interest like in the SHOE project, in the KA2 project, in the Knowledge Sharing Effort public Library (KSEL)... In a same way, we can find general ontologies in WWW's indexers like AltaVista or Yahoo! (Labrou and Finin, 1999). Second, a subset of a thesaurus can be used and can be extended to build ontologies. For (Martin, 1995), such a thesaurus is a wide linguistic ontology. Users should rarely have to add intermediate types but rather specialized precise types of WordNet in order to express the shades of meanings needed for the application (the disambiguation process described in the next section is then easiest). Finally, we can use tools (manual, semi-automatic or automatic) to build ontologies according to a set of typical data (KA2 project).

3.2 Disambiguating label of concepts

According to Knowledge Engineering researches, a concept is considered unambiguous and unique. However, in usual ontologies, a single label represents each concept. This choice is valid for the most part of IA processes but too superficial for linguistic based ones. Indeed, in our context a concept label is used as natural language entity which is a term. However, such a term is basically ambiguous. For example, if we find the term "chair" as a

concept label, either this is the "armchair" or this is the faculty member of a University. Therefore, we decide to disambiguate labels used to represent concepts of ontologies using hierarchical heuristics based on the ontology and the WordNet thesaurus (Miller, 1990). However, after our process, some labels of concepts may still have more than one related sense (generally between two and four). Since the ontology represents often knowledge of a specialized domain, in most cases the process cannot be completely automatic. Therefore, the user has to decide what sense to take when it's necessary.

A thesaurus can be viewed as a linguistic ontology (Guarino *et al*., 1999; Borgo *et al*., 1997). In WordNet, a "concept", called *a sense*, is defined with a single set of synonyms, called *a synset*. Therefore, concept in a thesaurus is unambiguous. For example, the first sense (sense 1) of the term "Person" is the synset {"individual#1", "person#1", "human#1", "mortal#1", "soul#2", "somebody#1", "someone#1"} (for each term the own sense number is given after the '#' symbol). We use WordNet because it is free available for research purpose and it is a broad coverage linguistic ontology (70 000 nodes). However, it does not include cross-part-of speech semantic relationship and it includes too much fine-grained sense-distinctions and lacks domain information (O'Hara *et al*., 1998) for text retrieval.

Like in the OntoSeek project (Guarino *et al*., 1999; Borgo *et al*., 1997), our approach adds linguistic attributes to classical ontologies using the WordNet thesaurus to improve our semi-automatic web site knowledge discovering. Guarino calls this process a *disambiguation process*. Thus, we obtain a terminologically oriented ontology (Martin, 1995). However, the handmade process OntoSeek uses ontologies not to define the knowledge of a site but to find user's data in a large classical database of web pages. Another project proposes a similar process: the Mikrokosmos project (O'Hara *et al*., 1998) to provide an knowledge-based machine translation process. This process is another semi-automatic process (the user can refined manually the resulting disambiguation). It studies several heuristics. The both major ones are a hierarchical match heuristic and a similarity heuristic. The hierarchical match heuristic matches the ontology hierarchy with the WordNet one according to the "isa" relationship. For (O'Hara *et al*., 1998), the hierarchical match heuristic seems to be the powerful one to select right senses. Therefore, we choose to use this heuristic and to improve it.

Our main goal is to associate to each concept of the ontology the right synset of the thesaurus. An ontology concept has a label, which is only one of its possible lexical forms. This lexical form helps the process to select corresponding synsets into the thesaurus (each term has often several meanings). Then, among these selected synsets, we only choose the more relevant ones. Accordingly, for each concept label of the ontology, the thesaurus provides several candidate synsets related to this concept. To select the relevant synset, we try to find if the synset context according to the hypernym relationship in the thesaurus is similar to the concept context according to the "isa" relationship in the ontology. During this process, we measure the *matching degree* between a synset and a concept of the ontology. It is evaluated taking the result of the matching process into account, namely the number of related concepts, the type of relationship, the depth of the different relationship... Therefore, our ontology is a terminologically oriented ontology (Martin, 1999) to ease rapid and simple knowledge representation, management, and use.

To highlight the disambiguation process, we apply it to the Thoth's University ontology. This ontology is a modification of the SHOE's one improves according our thesaurus (see results in section 5 to compare them). An ontology can be viewed as a set of paths according to the "isa" relationship. An isa-path is a list of concepts used to go from a root of the ontology (we suppose the ontology is an acyclic directed graph) to a leaf. To continue, some examples of isa-paths coming from the Thoth's University ontology:

- (Assistant Employee Person Organism Entity)
- (AssociateProfessor Professor Educator Person Organism Entity)
- (AssociateProfessor Professor FacultyMember Employee Person Organism Entity)
- (Chair Employee Person Organism Entity)
- (Chair Leader Person Organism Entity)
- (Chair Professor Educator Person Organism Entity)
- (Chair Professor FacultyMember Employee Person Organism Entity)

- (FullProfessor Professor Educator Person Organism Entity)
- (FullProfessor Professor FacultyMember Employee Person Organism Entity)
- (Lecturer Educator Person Organism Entity)
- (Lecturer FacultyMember Employee Person Organism Entity)
- (VisitingProfessor Professor Educator Person Organism Entity)
- (VisitingProfessor Professor FacultyMember Employee Person Organism Entity)...

In this example, we only take concept label having at least one sense in WordNet. Indeed, regarding unknown concepts, which are not in the thesaurus, we create a new sense (a technical sense) having the number 0. To search for synsets in WordNet corresponding to a concept, first we use the label and we translate it into a term. For example, "FacultyMember" is translated into "faculty member". If this search failed, we use the short description of the concept. For example, the concept "EmailAddress" is found using is short description "email".

In this ontology, we can find labels with multiple senses in WordNet (like "Chair" has 4 senses, "Director" has 4 senses, "Dean" has 3 senses...), others with only one (like "GraduatStudent" or "AssistantProfessor"), and then, some of them may have no sense (like "AdministrativeStaff"). To disambiguate label of concepts having multiple senses, we build hypernym paths from WordNet related to all candidate senses and we match them with isa-paths. A hypernym path is built using thesaurus hypernym relationships (the relationship between terms in the thesaurus equivalents to the isa relationship between concepts in the ontology) and candidate label senses (the number just after the term). For example:

- (Assistant 1 Person 1 Organism 1 Entity 1)
- (AssociateProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)
- (Chair 1 Artefact 1 Entity 1)
- (Chair 2 Work 3 Activity 1 HumanActivity 1)
- (Chair 3 Leader 1 Person 1 Organism 1 Entity 1)
- (Chair 4 Artefact 1 Entity 1)
- (FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)
- (Lecturer 1 Educator 1 Person 1 Organism 1 Entity 1)
- (Lecturer 2 Person 1 Organism 1 Entity 1)
- (Person 2 Entity 1)
- (VisitingProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)...

Then, we match "isa" paths (isa-paths) and hypernym paths (hypernym-paths) to find candidate senses for each concept label. To perform this matching process, we define five heuristics taking both the isa path and hypernym paths into account. For each heuristic, we associate a *matching degree* corresponding to the degree of the matching convenience the heuristic gives. The first heuristic is: only hypernym-paths having more than one term common to the isa path are computed. The second heuristic gives 0.95 to each sense (and consequently to corresponding synset) of the hypernym-path if the isa-path is fully included in this hypernym-path. This rule is the best case we can find. This gives results of table 1. In such a matching process, each concept label of an isa-path has the corresponding sense and a coefficient equals to 0.95. For example, "lecturer" sense 1 is a sense for the label "Lecturer" with the matching degree equals to 0.95.

The third heuristic gives 0.75 to each sense (and consequently to corresponding synset) of the hypernym-path if this path is fully included in the isa-path and if at least three terms are in both paths. This gives results of

Hypernym-path	isa-path	matching degree
(Chair 3 Leader 1 Person 1 Organism 1 Entity 1)	(Chair Leader Person Organism Entity)	0.95
(Dean 1 Chief 1 Leader 1 Person 1 Organism 1 Entity 1)	(Dean Chief Leader Person Organism Entity)	0.95
(Director 1 Chief 1 Leader 1 Person 1 Organism 1 Entity 1)	(Director Chief Leader Person Organism Entity)	0.95
(FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(FullProfessor Professor Educator Person Organism Entity)	0.95
(Lecturer 1 Educator 1 Person 1 Organism 1 Entity 1)	(Lecturer Educator Person Organism Entity)	0.95
(VisitingProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(VisitingProfessor Professor Educator Person Organism Entity)	0.95

Table 1: An extract of results for the second heuristic

Hypernym-path	isa-path	matching degree
(Assistant 1 Person 1 Organism 1 Entity 1)	(Assistant Employee Person Organism Entity)	0.75
(Learner 2 Person 1 Organism 1 Entity 1)	(GraduateStudent Student Learner Person Organism Entity)	0.75
(Lecturer 2 Person 1 Organism 1 Entity 1)	(Lecturer Educator Person Organism Entity)	0.75
(Lecturer 2 Person 1 Organism 1 Entity 1)	(Lecturer FacultyMember Employee Person Organism Entity)	0.75

Table 2: An extract of results for the third heuristic

table 2. Each concept label of the isa path which corresponds to a term in the hypernym-path has the sense of the corresponding hypernym and the coefficient 0.75. For example, "lecturer" sense 2 is a sense for the label "Lecturer" with the matching degree equals to 0.75.

The fourth heuristic gives 0.50 to senses of the hypernym-path if it exists labels of the isa-path in the hypernym path. The coefficient depends on the rate of isa-path labels included in the hypernym-path according to all isa-path labels, that is: $(0.50 * Length(Intersection(isa-path, hypernym-path)/Length(isa-path))$. This gives results of table 3. In the first example, the isa-path contains four labels having a sense in the hypernym-path. So the matching degree of "Leader 1" is $0.5 * 4/6 = 0.33$.

Hypernym-path	isa-path	matching degree
(Chair 3 Leader 1 Person 1 Organism 1 Entity 1)	(Director Chief Leader Person Organism Entity)	0.333
(FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(FullProfessor Professor FacultyMember Employee Person Organism Entity)	0.429
(FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(Lecturer Educator Person Organism Entity)	0.400
(FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(Dean Professor FacultyMember Employee Person Organism Entity)	0.347
(FullProfessor 1 Professor 1 FacultyMember 1 Educator 1 Person 1 Organism 1 Entity 1)	(Director Chief Leader Person Organism Entity)	0.250

Table 3: An extract of results for the fourth heuristic

The last heuristic gives 0.25 to senses of the hypernym-path if it exists terms of the hypernym-path in the isa-path. The coefficient depends on the rate of hypernym-path terms included in the isa-path according to all hypernym-path terms, that is: $(0.25 * Length(Intersection(isa-path, hypernym-path)/Length(hypernym-path))$. This gives results of table 4.

Hypernym-path	isa-path	matching degree
(Person 2 Entity 1)	(FullProfessor Professor FacultyMember Employee Person Organism Entity)	0.25
(Journal 5 Artefact 1 Entity 1)	(Journal Periodical Publication Creation Artefact Entity)	0.25
(Chair 1 Artefact 1 Entity 1)	(Magazine Periodical Publication Creation Artefact Entity)	0.167
(Chair 4 Artefact 1 Entity 1)	(Chair Professor Educator Person Organism Entity)	0.167
(University 2 Artefact 1 Entity 1)	(Magazine Periodical Publication Creation Artefact Entity)	0.167

Table 4: An extract of results for the fifth heuristic

For each sense, we provide its number (0 for unknown ones) and a matching degree depending on the heuristic used to find it. 1.0 is assigned to concept labels with unknown or single selected sense. Only those having a matching degree greater than zero are taken into account. If a sense is obtained by two heuristics, then the higher matching degree is taken into account. For example, in matches previously presented, we saw "FullProfessor 1" having 0.95, 0.75, 0.4... As a result, the matching degree for the sense 1 of the label (and concept) "FullProfessor" is 0.95. Therefore, since this is the only sense, its degree is modified to 1.0. Then, after this disambiguation process, we obtain a XML file. An extract of this file is shown in figure 2 (the matching degree is the "convenience" attribute).

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE ontology SYSTEM "file://localhost/../JipOnto.dtd">
<ontology id="university-ont" version="3.0" description="">
  <def-category name="Lecturer" short="lecturer" description="" sense="1" isa="Educator FacultyMember">
    <sense name="Lecturer" no="2" origin="WN" definition="someone who lectures professionally" convenience="0.75">
      <synset>lecturer#2</synset>
    </sense>
    <sense name="Lecturer" no="1" origin="WN" definition="a public lecturer at certain universities"
convenience="0.95">
      <synset>lector#1,lecturer#1,reader#4</synset>
    </sense>
  </def-category>
  <def-category name="Chair" short="chair" description="" sense="3" isa="Administrative Staff Professor Leader">
    <sense name="Chair" no="4" origin="WN" definition="an instrument of death by electrocution that resembles a chair"
convenience="0.2">
      <synset>chair#4,death chair#1,electric chair#1,hot seat#1</synset>
    </sense>
    <sense name="Chair" no="1" origin="WN" definition="a seat for one person, with a support for the back"
convenience="0.2">
      <synset>chair#1</synset>
    </sense>
    <sense name="Chair" no="3" origin="WN" definition="the officer who presides at the meetings of an organization"
convenience="0.95">
      <synset>chair#3,chairman#1,chairperson#1,chairwoman#1,president#4</synset>
    </sense>
  </def-category>
  <def-category name="Employee" short="employee" description="" sense="1" isa="Person">
    <sense name="Employee" no="1" origin="WN" definition="a worker who is hired to perform a job"
convenience="1.0">
      <synset>employee#1</synset>
    </sense>
  </def-category>...
</ontology>
```

Figure 2: An ontology after the automatic step of the disambiguation

In order to improve the matching process and to suppress ambiguity concerning concept label (which can be

matching with more than one synset), we can also remove from a concept of the ontology, senses (synsets) having a higher matching coefficient than another one in the same ontology. Finally, the user can select his/her sense for concept labels having multiple senses. To help him/her, we propose the WordNet definition for each selected sense and candidate senses are sorted according to the coefficient of matching convenience.

Remark: This entire process (and the final matching in section 5) is implemented using the Java Expert System Shell called Jess (Friedman-Hil, 2000), which used a CLIPS like language. Jess is a portable, extensible, fast reasoning engine written in Sun Microsystem's Java language. It was developed at Sandia National Laboratories (USA) and is distributed free of charge for academic use. Jess has a large, worldwide user community. Jess is commonly used in agent research, as it provides a convenient way to integrate complex reasoning capabilities into Java-based software. Jess users have developed many software extensions, including ones for fuzzy logic, database access, blackboards, and language understanding. Jess has been used to deploy many real systems, including some based on a multi-agent paradigm.

3.3 Results

We applied this disambiguation process to two versions of the University ontology:

1. the SHOE University ontology (74 concepts, 195 senses i.e. 2.26 sense by concepts),
2. the Thoth's University ontology (79 concepts, 199 sense i.e. 2.52 senses by concepts) see Appendix 2.

To evaluate this process, we classify concepts according to the level of disambiguation. The table 5 shows the rate of concepts concerned by each level. The first level concerns concepts having their right label sense not selected. The second level concerns concepts having their right label sense selected. The third level concerns concepts having the right label sense, which belongs to the best ones considering the matching degree. The fourth level concerns concepts, which have the right label sense with the unique best matching degree. Finally, a concept in the fifth level is a concept having only their right label sense selected.

Level	The right label sense...	SHOE's University ontology	Thoth's University ontology
0	...is not selected.	1.35%	0%
1	...is selected.	98.65%	100%
2	...belongs the best ones.	97.30%	100%
3	...is the best one.	74.32%	96.20%
4	...is the only selected sense.	62.16%	67.09%

Table 5: Results of the disambiguation process

Regarding these results, even if the disambiguation process is not complete for a given concept, the user has a good chance to find the right label sense first (or among the first ones). This chance is increased if the ontology is convenient for the thesaurus. Indeed, the Thoth's ontology is built according to a substantial modification of the SHOE's one according to WordNet. As a result, the disambiguation process is more efficient on the Thoth's ontology than the SHOE's one.

Table 6 gives several samples of concepts for each level. In this table, a concept is described by its label, its right label sense (after the '#' sign), the number of candidate label senses at the end of the disambiguation process in comparison of the number of all its possible label senses, and a list of each label sense number and its matching degree.

We can see the SHOE's University ontology is not really convenient regarding the disambiguation process with WordNet. Therefore, we decided to modify it to improve our process. For example, the concept "Faculty" had a wrong label because it describes the function "faculty member" (its short description) and not the organism. Hence, we called it "FacultyMember". This ontology had also inconsistencies with WordNet regarding the isa-path "(Program Organization SocialGroup Agent Entity)". In WordNet, a program is a document but not an organization. Next, this ontology has only one root "Entity". Conversely, the WordNet hierarchy has multiple roots depending class of concepts. For example, a "publication" is not an "entity" but an "artefact"... Another

Level	SHOE's University ontology	Thoth's University ontology
0	Agent #1 - 4/5 (5/0.3 2/0.3 4/0.3 3/0.25)	Empty
1	Document #1 - 4/4 (4/0.25 1/0.17 3/0.25 2/0.17)	Thesis #2 - 2/2 (1/0.38 2/0.38) School #1 - 3/7 (6/0.33 4/0.95 1/0.95)
2	Dean #1 - 3/3 (3/0.75 1/0.75 2/0.75)	The same as level 1
3	Work #1 - 5/7 (3/0.25 1/0.38 5/0.25 6/0.25 2/0.25)	Magazine #2 - 5/6 (6/.25 4/.25 1/.25 5/.25 2/.75) Information #1 - 2/5 (1/0.95 5/0.38)
4	Publication #1 - 1/3 (1/0.38) EmailAddress #1 - 1/1 (1/1.0)	Professor #1 - 1/1 (1/1.0) Address #2 - 1/7 (2/0.95)

Table 6: Samples of concepts in each level of disambiguation

major problem, mainly shown by the matching process, concerns the relative position of "Agent" and "Person". This ontology gives "(Person Agent Entity)". On the opposite, WordNet gives "(Agent 2 Person 1 Organism 1 Entity 1)". Adding to other problems concerning the "isa" relation, we decided to build the Thoth's University ontology. This ontology uses a wide part of the SHOE's one but its organization is rather different.

Consequently, this process allows the user to verify and to correct the ontology too. Looking to the matching result help us to correct the ontology according to the referring thesaurus. Table 5 shows that manipulating ontology can improve the disambiguation process. These manipulations can be: adding or removing concepts, changing concepts labels or descriptions...

4 Index building

In Information Retrieval community, it is assumed that discriminating word senses in requests and documents significantly improves the performance of information retrieval systems. Some experiments were lead (Gonzalo *et al*., 1998) on documents and requests semantically hand-annotated: the precision was greatly enhanced. We saw in section 3 that manual annotation of web pages with ontology is a titanic task. Therefore, we propose a semi-automatic indexer of web pages. The major problem is to semi-automatically determine the specific content of documents. The natural language processing community commonly admitted that terms involved in text, carry relevant information on text content. Several works (Bourigault, 1994; Daille, 1994) underline that terms pattern and frequency are determinant to extract terms from large corpora from a single domain. Unfortunately, web pages are often small and web sites often cover multiple domains. However, web pages are structured according to HTML markers, which can be used for weighting terms importance in pages and then can be exploited for characterizing web page content.

4.1 Terms extraction

The term extraction begins by (1) removing HTML marker from web pages, (2) dividing the text in independent sentences, (3) lemmatizing words included in the page. Next, web pages are annotated with part of speech tags using the Brill tagger (Brill, 1995). As a result, each word in a page is annotated with its corresponding grammatical category (noun, modifier...). Finally, the surface structure of sentences is then analyzed using term patterns to provide well-formed terms. Several term patterns are given in table 7.

To each well-formed term is assigned a coefficient C according to the term frequency and the weight of HTML Markers. This last coefficient is called the *weighted frequency*. For example, the "TITLE" marker weights by 10, "KEYWORD" by 9, "H1" by 3... (see table 8)

In a web page containing m different terms, for a given term T, the $C(T)$ coefficient is determine as the sum for the n occurrences of the term T of their associated HTML marker weight. The result is then normalized. This calculus is shown in formula 1 where $htmlcoefficient_i(T)$ corresponds to the HTML marker weight associated with the ith occurrence of the term T.

Term patterns
Noun
Noun Noun
Noun "of" Noun
Adjective Noun
...

Table 7: Samples of term patterns

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 8: Higher coefficients associated with HTML markers

$$C(T_k) = \frac{\sum_{i=1}^{n_j} htmlcoefficient_i(T_j)}{max_{k\in[1,m]} \sum_{i=1}^{n_k} htmlcoefficient_i(T_k)} \quad (1)$$

Table 9 shows some results extracted from an experiment on the web site “http://www.cs.washington.edu/”. This web page is the home-page of the department of computer science of the University of Washington. A coefficient equal to 1.0 means the term (here “Washington”) is the most relevant index. Other term coefficients are calculated according to this one. We have extracted well-formed terms like single term (“science”...) and complex term (“university washington”...). Note that the term “university washington” provides another term “university of washington”. We provide several forms for a term to improve the term retrieval process in the thesaurus.

4.2 Page concept determination

During the term extracting process, well-formed terms and their coefficient were respectively extracted and calculated. The well-formed terms are forms representing a particular concept. Different forms may represent a same concept (i.e. chair, professorship). In order to determine not only the term set included in a page but also the concept set included in a page, linguistic ontologies are used. A linguistic ontology can be viewed as a machine readable dictionary structured around groups of synonymous words, which represent a concept. Moreover, it provides explicit relationships among group of synonymous (i.e. hypernym relationship, meronym relationship...). Linguistic ontology allows the mapping of concepts and words or terms. Therefor, in our context, linguistic ontology is used to generate all concepts corresponding to well-formed term. In our experiment, we have still used the WordNet linguistic ontology (Miller, 1990).

Weighted coefficient	Terms
1.00	washington
0.83	science
0.67	university washington
0.67	university of washington
0.67	university
0.67	engineering
0.50	program
0.50	computer science
0.50	computer
0.50	college
0.33	uw
0.33	student
0.33	seattle
0.33	member
0.33	major
0.33	field
...	...

Table 9: Terms extracted from a web page

The process to generate candidate synsets is quite simple: from all extracted terms, all candidate concepts (all senses) are generated using WordNet. This thesaurus is a broad coverage linguistic ontology but it does not cover all the terms included in web page. If a term does not exist in WordNet, a specific sense is generated. Then, a convenience coefficient is calculated using a semantic similarity measure (Figure 3). It measures for a given term (a form), the convenience with all possible concepts it could represent. The calculus takes terms context into account (the page in which it appears and its neighborhood).

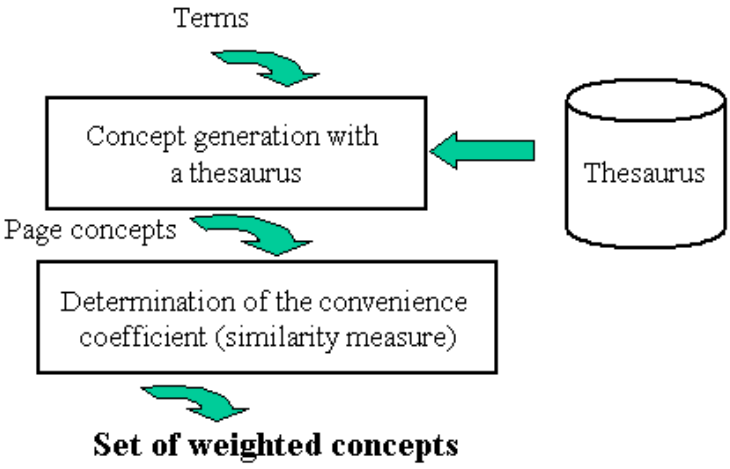

Figure 3: Concept generating and weighting

Although, many measures of similarity are defined in related works. For (Lee *et al*, 1993), 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

being 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 (Resnik, 1999) with this approach is that it relies on the notion that links in the taxonomy represent uniform distances, which is most of the time a wrong assumption. (Resnik, 1999) describes an alternative way to evaluate semantic similarity in a taxonomy based of the notion of information content. All links in a taxonomy are weighted with an estimated probability (concept occurrence in corpora), which measures the information content of a concept. The main idea is: the more information two concepts share, the 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 witch is equal to $-log(P(c))$.

Some authors (O'Hara *et al*, 1998) exploit other information type to determine the own sense of a term during the disambiguation of ontology concept label. They exploit term definitions in a thesaurus and the Resnik similarity measure in order to disambiguate ontology concept labels. After empirical evaluation, they conclude that the term definition is not a major criteria for such a problem.

In our context, we take another approach between the simple edge-counting approach and the Resnik's approach to determine semantic similarity. (Wu *et al*., 1994) propose a similarity measure related to the edge distances 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 2 where c is the most specific subsumer of c_1 and c_2, $depth(c)$ is the edge number from c to the taxonomy root, and $depth_c(c_i)$ with i in $\{1, 2\}$ is the edge number from c_i to the taxonomy root through c.

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

This measure performs a little worse than the Resnik's measure but better than the traditional edge-counting measure (Resnik, 1999). However, Resnik estimates the concept probability using noun frequencies from the Brown Corpus of American English, which is not convenient in our context (web pages). Indeed, for specific and technical domain such corpora are not available. Evaluating the coefficient with general corpora as the Brown Corpus can produce wrong result for a specific and specialized domain.

The convenience coefficient, for a specific concept, is the normalized sum of all semantic similarities calculated with all other concepts included in the studied web pages. In this formula, a specific concept is unified with the corresponding synset in WordNet. The measure is shown in formulas 3 and 4, where a term T_k has l_k associated synsets, and there are m terms in the studied web pages.

$$simsum(synset_i(T_k)) = \sum_{j \in [1,k-1] \cup [k+1,m]} \sum_{l=1}^{l_j} sim(synset_i(T_k), synset_l(T_j)) \tag{3}$$

$$conv(synset_i(T_k)) = \frac{simsum(synset_i(T_k))}{max_{j \in [1,l_k]} simsum(synset_j(T_k))} \tag{4}$$

A bias of 0.05 is applied on results in order to avoid a convenience coefficient equal to 1.0 if there is more than one candidate sense. Moreover, in order to discriminate results, only similarities between concepts greater than alpha (alpha is an experimental threshold) are hold. If only one synset is associated with a term, its convenience coefficient is valuated to 1.0 (it means that the term is associated with only one synset in the thesaurus or it is not present in the thesaurus). Then, figure 4 shows an extract of the XML file containing candidate concepts (the DTD is presented in appendix 1).

5 Associating concepts and synsets

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE indexpage SYSTEM "file://localhost/…/ontologies/index-page.dtd">
<indexpage>
<nindex page="http://www.cs.washington.edu/" convenience="1.00" frequence="0.67">university of
washington#0</nindex>
<index page="http://www.cs.washington.edu/" convenience="0.95" frequence="1.00">science#1,scientific
knowledge#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.95" frequence="1.00">science#2,scientific
discipline#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.95" frequence="1.00">skill#2,science#3</index>
<nindex page="http://www.cs.washington.edu/" convenience="1.00" frequence="0.67">university washington#0</nindex>
<index page="http://www.cs.washington.edu/" convenience="1.00" frequence="0.50">computer science#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.1" frequence="1.00">washington#1,american
capital#1,capital of the united
states#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.7" frequence="1.00">washington#2,evergreen
state#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.3" frequence="1.00">capitol#1,washington#3</index>
<index page="http://www.cs.washington.edu/" convenience="0.3" frequence="1.00">washington#4,george
washington#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.3" frequence="0.67">university#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.5" frequence="0.67">university#2</index>
<index page="http://www.cs.washington.edu/" convenience="0.7" frequence="0.67">university#3</index>
<index page="http://www.cs.washington.edu/" convenience="0.95" frequence="0.50">program#3,programme#3,computer
program#1,computer  programme#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.95" frequence="0.50">course of
study#1,program#4,curriculum#1,syllabus#1</index>
<index page="http://www.cs.washington.edu/" convenience="0.5"
frequence="0.50">broadcast#2,program#5,programme#2</index>
<index page="http://www.cs.washington.edu/" convenience="0.1" frequence="0.50">program#6,programme#6</index>
<index page="http://www.cs.washington.edu/" convenience="0.1" frequence="0.50">platform#2,political
platform#1,political
program#1,program#7</index>
<index page="http://www.cs.washington.edu/" convenience="0.5" frequence="0.50">program#8,programme#1</index>…
</indexpage>
```

Figure 4: Extract of a generated index

5.1 the process

At this point, we have on the one hand an ontology for which concept labels are disambiguated and on the other hand possible senses in each HTML pages of the Web site with their relative frequency, and their evaluated convenience. In the next step, indexes are matched with concepts of ontologies. For each sense, we search for the same one in the index. If it exists, concerned web pages and coefficients are added to it.

For the moment, our process gets pages containing concepts of the ontology. However, it does not take the weighted frequency of synsets into account. Consequently, a concept that appears only one time in a page allows this page to be referred by the ontology. For this reason, we added a *frequency threshold* to consider a concept only if its weighted frequency (section 4.1) is greater or equal than this threshold. In the next section, we will present several indexation processes according to the evolution of the threshold.

To evaluate the appropriateness of an ontology according to a of HTML pages, four typical coefficients are calculated:

1. the rate of pages concerned with its concepts, called *the covering degree*, which gives the number of web pages that involve at least one concept of the ontology,

2. the rate of its concepts directly involved in HTML pages, called *the direct indexing degree*,

3. the rate of its concepts involved (directly or by the way of more specialized concepts), called *the global indexing degree*,

4. the average convenience degree of candidate concepts from pages selected by the ontology.

Finally, a relevant ontology is an ontology having these coefficients close to 1.0. A high covering degree implies a wide proportion of the pages contain concepts of the ontology. A high direct indexing degree implies a lot of concepts can be found in the pages. A high value for this couple of coefficient is quite important. Namely, we can have a site where only one page contains the ontology (this gives an indexing degree at 1 and a little covering degree). In the same way, all pages can contain a general concept like "Entity" in the head of each page (this gives a weak indexing degree but a covering degree equals to 1).

The indexation process can also highlight indexes, which do not match with concepts of ontologies. In this case, we may search for ontologies related to this index. In the future, one can redo the indexation process either when the site content notably evolves or when the used ontologies are updated. This process can only be executed with modified pages.

5.2 Some results of the general process

To evaluate the performance of our indexation process, we tried it with the Thoth's University ontology and three web sites:

1. "http://www.cs.washington.edu/": the University of Washington Department of Computer Science & Engineering web site (we will call it *Washington*), which has 1 315 pages and 480 135 candidate concepts (among these concepts only between 250 and 7 810 are selected related to the frequency threshold).

2. "http://www.cookingwithkids.com/": the web site dedicated to the book "Cooking with Kids for Dummies" by Kate Heyhoe was published in March, 1999 by IDG Books (we will call it *Cooking with Kids*), which has 100 pages and 39 426 candidate concepts (among these concepts only between 2 and 212 are selected related to the frequency threshold).

3. "http://www.sofaandchair.com/": the web site of "get FURNISHED", a virtual online home decor store (we will call it *Sofa and Chair*), which has 68 pages and 19 597 candidate concepts (among these concepts only between 2 and 136 are selected related to the frequency threshold).

Washington was chosen for its *a priori* high relevance with the Thoth's University ontology. In the same way, the two other sites was chosen for their *a priori* weak relevance with this ontology. For each web site, we have processed our indexation with the frequency threshold that goes up from 0 (all senses are accepted in each pages) to 1 (only the most important ones are accepted in each pages). For each process, we turn ours attention to the covering degree, the direct indexation degree, and the global indexation degree. Figures 5, and 6 shows results for *Washington*, *Cooking with Kids* and *Sofa and Chair*. Appendix 3 shows an extract of the resulting XML file for the *Washington* web site with the frequency threshold equal to 0.5.

Figure 5 shows the evolution of the covering degree (the rate of selected pages) when the frequency threshold goes up. This graph brings out the covering degree of *Washington* pages has a slight and steady decrease. However, one may observe there is a substantial fall up to 0.3 for *Cooking with Kids* mainly and in some degrees for *Sofa and Chair*. It must be pointed out the strange beginning of the *Cooking with Kids* curve: 99% of pages are indexed with the Thoth's University ontology using a frequency threshold equals to zero. It can be accounted by the subject of this site: It talks about a "book", a concept of this ontology. Each page contains this concept with a rather weak frequency. Therefore, when the threshold goes up, the number of pages containing this concept quickly falls and the rate plummets. In this connection, the covering degree of the *Sofa and Chair* web site is rather low but the begining is also important. The responsible concept is "Chair". The University concept is also present but with a weak convenience degree.

Figure 6 shows the evolution of both the direct indexation degree (number of concepts directed involved in pages) of each web site. In the same way, the *Washington* web site has a rather steady curve and rather higher than the others. One can be surprised by a too low curve. This can be explained (1) by the presence of numerous personal home-pages, (2) by an ontology a little bit to general, and (3) by a site which concerns a computer science research laboratory, and not describes all University capabilities. Likewise to the covering degree, curves of the two other sites are rather important in the beginning. This can be attributed on the one hand to "Chair", "Information", "Work" for the first one and on the other hand to "Book", "Course", "Work" and "Person" for the second one.

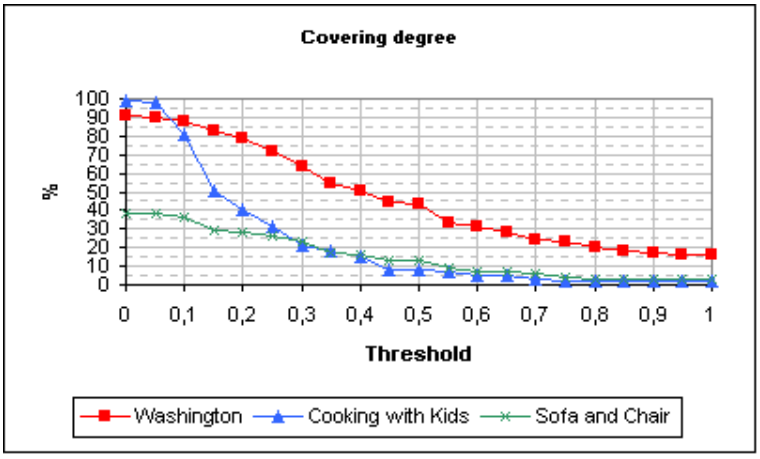

Figure 5: The covering degree analysis

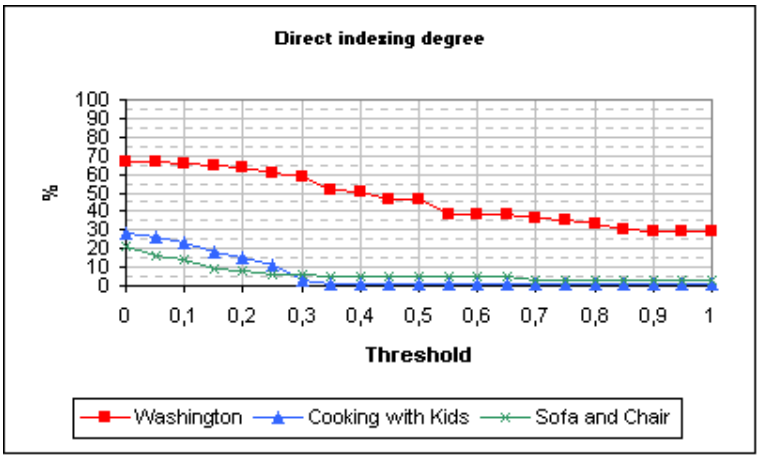

Figure 6: The direct indexation degree analysis

6 Conclusions

In this paper, we have presented a semi-automatic process to index a web site by its content. On the one hand, we presented a process to disambiguate concept label of ontologies, and on the other hand a process to extract, with natural language features, candidate concepts from the web pages. Then, we presented the matching process between ontology concepts and candidate concepts included in web pages. Experiments lead on three web sites (one linked to the ontology domain, two others linked to other domain) have shown the appropriateness of our approach. It is an efficient process assuming that we know the domain(s) covered by the web site pages and we use an ontology(ies) linked to this domain(s).

This Thoth's ontology-based indexation process will be used in the context of multi-agent system like Bonom (Cazalens *et al.*, 2000) where indexed information is not centralized but is distributed on web sites. Agents dedicated to a web site are able to index this site and to answer queries. Our indexation process can improve these tasks. Indeed, we give the capability to a web site agent (if its web-master has selected convenient ontologies) to determine its content according to these ontologies and then to answer to users queries, which could be automatically (or semi-automatically) disambiguated according to these same ontologies. Another field of application

will be competitive intelligence and information retrieval on invisible web. Indeed, only 55% of information on the web is pointed by search engines (http://www.strasbourg.cci.fr/infoEco/veille2.htm; CCI Strasbourg, France). Having domain specific ontologies and several web sites of this domain, we can propose a web agent which is able to follow hyperlinks from page to page, to index the encountered web sites and to select relevant ones according to ontologies. Among this set of relevant selected sites, some of them can belong to the invisible web.

7 Appendix

7.1 DTDs

7.1.1 Ontologies

```
<?xml version="1.0" encoding="UTF-8"?>

<!-- XML DTD for JIP ontologies -->
<!-- Last Mod: 27/10/2000 -->
<!-- Version 2.0 -->

<!ELEMENT ontology      (use-ontology | def-category | def-relation |
                         def-rename | def-inference | def-constant |
                         def-type)* >
<!ATTLIST ontology
        id                      CDATA   #REQUIRED
        version                 CDATA   #REQUIRED
        description             CDATA   #IMPLIED
        declarators             CDATA   #IMPLIED
        backward-compatible-with CDATA  #IMPLIED >

<!ELEMENT use-ontology  EMPTY>
<!ATTLIST use-ontology
        id              CDATA   #REQUIRED
        version         CDATA   #REQUIRED
        prefix          CDATA   #REQUIRED
        url             CDATA   #IMPLIED >

<!ELEMENT def-category  (sense*)>
<!ATTLIST def-category
        name            CDATA   #REQUIRED
        isa             CDATA   #IMPLIED
        description     CDATA   #IMPLIED
        short           CDATA   #IMPLIED
        sense           CDATA   "UNKNOWN" >

<!ELEMENT sense         (synset, page*)>
<!ATTLIST sense
        no              CDATA   #REQUIRED
        name            CDATA   #REQUIRED
        origin          CDATA   "WN"
        definition      CDATA   ""
        convenience     CDATA   #REQUIRED>

<!ELEMENT synset        (#PCDATA)>

<!ELEMENT page          EMPTY>
<!ATTLIST page
        name            CDATA   #REQUIRED
        frequence       CDATA   #REQUIRED
        convenience     CDATA   #REQUIRED>
...
```

7.1.2 Indexes

```
<?xml version="1.0" encoding="UTF-8"?>

<!-- XML DTD for JIP indexes -->
<!-- Last Mod: 28/10/2000 -->
<!-- Version 1.0 -->


<!ELEMENT indexpage (index | nindex)*>
<!ELEMENT index (#PCDATA)>
```

```
<!ATTLIST index         page            CDATA #REQUIRED
                        frequence       CDATA #REQUIRED
                        convenience     CDATA #REQUIRED>

<!ELEMENT nindex (#PCDATA)>
<!ATTLIST nindex        page            CDATA #REQUIRED
                        frequence       CDATA #REQUIRED
                        convenience     CDATA #REQUIRED>
```

7.2 Thot's University ontology

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>

<!DOCTYPE ontology SYSTEM "file://localhost/home/.../ontologies/JipOnto.dtd">

  <ontology id="university-ont" version="3.0" description="...">

    <def-category name="HumanActivity" sense="1"/>

    <def-category name="Activity" isa="HumanActivity" short="activity" sense="1"/>
    <def-category name="Work" isa="Activity" short="work" sense="1"/>
    <def-category name="Recreation" isa="Activity" short="recreation" sense="1"/>
    <def-category name="Process" isa="Activity" short="process" sense="1"/>
    <def-category name="Course" isa="Work" short="teaching course" sense="1"/>
    <def-category name="Research" isa="Work" short="research work" sense="1"/>

    <def-category name="WebResource" isa="Creation" short="web ressource" sense="0"/>
    <def-category name="Image" short="picture" isa="WebResource" sense="3"/>
    <def-category name="WebPage" isa="WebResource" short="html page" sense="0"/>

    <def-category name="SocialGroup" short="social group" sense="1"/>
    <def-category name="Organization" isa="SocialGroup" short="organization" sense="1"/>
    <def-category name="EducationOrganization" isa="Organization" short="education organization" sense="0"/>
    <def-category name="Department" isa="EducationOrganization" short="university department" sense="1"/>
    <def-category name="Institute" isa="EducationOrganization" short="institute" sense="1"/>
    <def-category name="School" isa="EducationOrganization" short="school" sense="1"/>
    <def-category name="ResearchGroup" isa="EducationOrganization" short="research group" sense="0" />
    <def-category name="University" isa="EducationOrganization" short="university" sense="3"/>

    <def-category name="Entity" short="entity" sense="1"/>
    <def-category name="Organism" isa="Entity" short="organism" sense="1"/>
    <def-category name="Person" isa="Organism" short="person" sense="1"/>
    <def-category name="Employee" isa="Person" short="employee" sense="1"/>
    <def-category name="FacultyMember" isa="Employee" short="faculty member" sense="1"/>
    <def-category name="Educator" isa="Person" sense="1"/>
    <def-category name="Professor" isa="FacultyMember Educator" short="professor" sense="1"/>
    <def-category name="AssistantProfessor" isa="Professor" short="assistant professor" sense="1"/>
    <def-category name="AssociateProfessor" isa="Professor" short="associate professor" sense="1"/>
    <def-category name="FullProfessor" isa="Professor" short="full professor" sense="1" />
    <def-category name="VisitingProfessor" isa="Professor" short="visiting professor" sense="1"/>
    <def-category name="Lecturer" isa="FacultyMember Educator" short="lecturer" sense="1"/>
    <def-category name="PostDoc" isa="FacultyMember" short="post-doctorate" sense="0"/>
    <def-category name="Assistant" isa="Employee" short="assistant" sense="1"/>
    <def-category name="ResearchAssistant" isa="Assistant" short="university research assistant" sense="0"/>
    <def-category name="TeachingAssistant" isa="Assistant Educator" short="university teaching assistant" sense="0"/>
    <def-category name="Leader" isa="Person" sense="1"/>
    <def-category name="Chief" isa="Leader" sense="1"/>
    <def-category name="AdministrativeStaff" isa="Employee" short="administrative staff worker" sense="0"/>
    <def-category name="Director" isa="AdministrativeStaff Chief" short="director" sense="1"/>
    <def-category name="Chair" isa="AdministrativeStaff Professor Leader" short="chair" sense="3"/>
    <def-category name="Dean" isa="AdministrativeStaff Professor Chief" short="dean" sense="1"/>
    <def-category name="ClericalStaff" isa="AdministrativeStaff" short="clerical staff worker" sense="0"/>
    <def-category name="SystemsStaff" isa="AdministrativeStaff" short="systems staff worker" sense="0"/>
    <def-category name="Learner" isa="Person" sense="1"/>
    <def-category name="Student" isa="Learner" short="student" sense="1"/>
    <def-category name="UndergraduateStudent" isa="Student" short="undergraduate student" sense="0"/>
    <def-category name="GraduateStudent" isa="Student" short="graduate student" sense="1"/>

    <def-category name="Artefact" isa="Entity" sense="1"/>
    <def-category name="Creation" isa="Artefact" sense="2"/>
    <def-category name="Publication" isa="Creation Communication" sense="1" />
    <def-category name="Article" isa="Creation" sense="1"/>
    <def-category name="Book" isa="Publication" sense="1"/>
    <def-category name="BookArticle" isa="Article" sense="0"/>
    <def-category name="ConferencePaper" isa="Article" sense="0"/>
    <def-category name="Thesis" isa="Publication" sense="2"/>
```

```
    <def-category name="DoctoralThesis" isa="Thesis" short="phd thesis" sense="0"/>
    <def-category name="Periodical" isa="Publication" sense="1"/>
    <def-category name="Journal" isa="Periodical" sense="2"/>
    <def-category name="JournalArticle" isa="Article" sense="0"/>
    <def-category name="Magazine" isa="Periodical" sense="2"/>
    <def-category name="MastersThesis" isa="Thesis" sense="0"/>
    <def-category name="Proceedings" isa="Publication" sense="2"/>
    <def-category name="WorkshopPaper" isa="Article" sense="0"/>

    <def-category name="Location" isa="Entity" sense="1"/>
    <def-category name="Address" short="address"  isa= "Location" sense="2"/>
    <def-category name="StreetAddress" isa="Address" sense="1" />
    <def-category name="CityAddress" isa="Address" sense="0"/>
    <def-category name="StateAddress" isa="Address" sense="0"/>
    <def-category name="ZipAddress" isa="Address" sense="0"/>
    <def-category name="EmailAddress" short="email" isa="Address" sense="1"/>
    <def-category name="HomePhone" isa="Address" short="home phone number" sense="0"/>
    <def-category name="WorkPhone" isa="Address" short="work phone number" sense="0"/>
    <def-category name="OrgPhone" isa="Address" short="organization phone number" sense="0"/>

    <def-category name="SocialRelation" short="social relation" sense="1"/>
    <def-category name="Communication" isa="SocialRelation" sense="2"/>
    <def-category name="Information" isa="Communication" sense="1"/>
    <def-category name="Program" isa="Information" short="program" sense="4"/>

    <def-category name="Schedule" isa="Communication" short="schedule" sense="2"/>
    <def-category name="Meeting" isa="SocialGroup" sense="1"/>
    <def-category name="Conference" isa="Meeting" short="conference" sense="1"/>

  </ontology>
```

7.3 Extract of an indexation process

Site : “http://www.cs.washington.edu/”

Frequency threshold : 0.3

Ontology : completely disambiguated Thoth’s University’s ontology

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE ontology SYSTEM "file://localhost/home/.../ontologies/JipOnto.dtd">
  <def-category name="Lecturer" short="lecturer" description="" sense="1" isa="FacultyMember Educator">
    <sense name="Lecturer" no="1" origin="wn" convenience="1.0">
      <synset>lector#1,lecturer#1,reader#4</synset>
      <page name="http://www.cs.washington.edu/htbin-post/mvis/mvis/TVtalks" frequence="0.5" convenience="0.45"/>
      <page name="http://www.cs.washington.edu/people/faculty/karp.html" frequence="0.33" convenience="0.85"/>
      <page name="http://www.cs.washington.edu/lab/sieg/labs/cse-card-access.html" frequence="0.3" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/news/1999DLS.html" frequence="0.75" convenience="0.0"/>
      <page name="http://www.cs.washington.edu/news/2000DLS.html" frequence="0.75" convenience="0.45"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/dls/" frequence="1.0" convenience="0.0"/>
      <page name="http://www.cs.washington.edu/homes/dickey/grad-brochure-blurb.htm" frequence="0.5" convenience="0.75"/>
    </sense>
  </def-category>
  <def-category name="VisitingProfessor" short="visiting professor" description="" sense="1" isa="Professor">
    <sense name="VisitingProfessor" no="1" origin="wn" convenience="1.0">
      <synset>visiting professor#1</synset>
    </sense>
  </def-category>
  <def-category name="Employee" short="employee" description="" sense="1" isa="Person">
    <sense name="Employee" no="1" origin="wn" convenience="1.0">
      <synset>employee#1</synset>
      <page name="http://www.cs.washington.edu/homes/lazowska/impact/statsci.html" frequence="0.36" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/affiliates/" frequence="0.43" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/chair/telecommuting.html" frequence="1.0" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/chair/boards.html" frequence="0.43" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/ott/CSE_affiliate_edl.htm" frequence="0.53" convenience="1.0"/>
    </sense>
  </def-category>
  <def-category name="Chair" short="chair" description="" sense="3" isa="Leader AdministrativeStaff Professor">
    <sense name="Chair" no="3" origin="wn" convenience="1.0">
      <synset>chair#3,chairman#1,chairperson#1,chairwoman#1,president#4</synset>
      <page name="http://www.cs.washington.edu/commercialization/lwop.html" frequence="0.38" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/people/staff/people_who_can_help.html" frequence="0.5" convenience="0.75"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/" frequence="0.63" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/ARL/committee.html" frequence="0.6" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/info/contact/" frequence="1.0" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/people/acm/people/" frequence="0.5" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/people/faculty/young.html" frequence="0.7" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/hightech/ht/tsld001.htm" frequence="0.5" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/press/stranger/diorio/" frequence="0.33" convenience="0.85"/>
      <page name="http://www.cs.washington.edu/commercialization/principles.html" frequence="0.3" convenience="0.75"/>
      <page name="http://www.cs.washington.edu/leadership/sld002.htm" frequence="1.0" convenience="0.55"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/karp/whitehouse.html" frequence="0.33" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/people/faculty/baer.html" frequence="0.43" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/people/faculty/lazowska/" frequence="0.63" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/chair/2000/2000.fac.htm" frequence="0.42" convenience="0.65"/>
      <page name="http://www.cs.washington.edu/lab/quotes.html" frequence="0.6" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/cra/case/" frequence="0.57" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/ip/" frequence="0.5" convenience="0.55"/>
```

```
      <page name="http://www.cs.washington.edu/homes/lazowska/ott/student.companies.html" frequence="0.33" convenience="0.55"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/lazowska.html" frequence="0.66" convenience="0.95"/>
      <page name="http://www.cs.washington.edu/people/faculty/tanimoto.html" frequence="0.9" convenience="1.0"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/chair/summer.support.html" frequence="0.33" convenience="0.55"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/hightech/ht/index.html" frequence="0.5" convenience="0.65"/>
    </sense>
  </def-category>
...
<ontology id="university-ont" version="3.0" description="">
```

8 References

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

R. Bayardo *et al*., "InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environment", In Proceedings of the ACM SIGMOD, ACM Press, 6, 1997, pp. 195-206.

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.

D. Bourigault, "Lexter, un logiciel d'extraction de terminologie: application à l'acquisition de connaissance à partir de textes", PHD Thesis at EHESS, Paris, 1994.

S. Borgo, N. Guarino, C. Masolo, and G. Vetere, "Using a Large Linguistic Ontology for Internet-Based Retrieval of Object-Oriented Components", In Proceedings of SEKE, Madrid, Spain, Jun. 18-20, 1997.

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.

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.

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

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.

E. J. Friedman-Hil and the Sandia Corporation. "Jess: the Java Expert System Shell". Distributed Computing Systems, Sandia National Laboratories, Livermore, CA, Version 6.02a, http://herzberg.ca.sandia.gov/jess/

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.

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, "Indexing with WordNet Synsets can Improve Text Retrieval", workshop on "usage of WordNet in natural language processing systems", Coling-ACL'98

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.

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.

V. Kashyap, and M. Rusinkiewicz, "Modeling and Querying Textual Data Using E-R Models and SQL", In Proceedings of Workshop on Management of Semi-Structured Data, 1997.

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.

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.

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.

H. Lieberman. "Letizia: An Agent That Assists Web Browsing". In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.

P. Martin, “Using the WordNet Concept Catalog and a Relation Hierarchy for Knowledge Acquisition”, In Proceedings of Peirce’95, Aug. 18, 1995.

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).

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

A. Moukas. “Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem”. In Proceedings of Practical Applications of Intelligent Agents and Multi-Agents Technology Conference, 1996.

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

M. Pazzani, J. Muramastu, and D. Billsus, “Syskill & Webert: Identifying Interesting Web Sites”, AAAI Spring Symposium on Machine Learning in Information Access, Stanford, CA, USA, 1996.

M. Nodine *et al*., “Active Information Gathering in InfoSleuth”, International Journal of Cooperative Information Systems, 5(1/2), 2000.

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.

M. Slodzian, “WordNet: what about its linguistic relevancy?”, Workshop Ontologies and Texts.

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

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

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

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.

Web Site Indexation and Ontologies

E. Desmontils & C. Jacquin

Abstract

This report presents a new approach to index a web site using ontologies and natural language techniques for Internet information retrieval. Ontologies are used to index a web site and, as a result to represent the web pages content. First, a linguistic ontology (a thesaurus) is used to disambiguate the label of ontology concepts. This disambiguation process uses several hierarchical heuristics that take advantage of the "isa" relationship on both the ontology and the thesaurus. Natural language techniques based on extraction of well-formed terms are also presented. They used a web page particularity: HTML markers. Then, from the previous extracted terms, associated concepts are generated using a linguistic ontology and a semantic similarity measure. At each candidate concept is associated a couple of coefficient: the convenience coefficient and the weighted frequency coefficient. The match between web page concepts and ontology concepts is then presented. Moreover, results about different web sites on different domains are presented. They point out the good accuracy of our ontology indexation process.

Categories and Subject Descriptors: H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Indexing methods, Linguistic processing, Thesauruses*; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Search process*

Additional Key Words and Phrases: World Wide Web, Information Retrieval, Internet, Indexing Web Pages, Ontologies, Semantic Indexation, Disambiguation Process

IRIN, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3