Ontology enrichment and indexing process

E. Desmontils, C. Jacquin, L. Simon

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

— Ingénierie des Connaissances —

RESEARCH REPORT

N° 03.05

Mai 2003

E. Desmontils, C. Jacquin, L. Simon
Ontology enrichment and indexing process
18 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`

Ontology enrichment and indexing process

E. Desmontils, C. Jacquin, L. Simon
{desmontils, jacquin, simon}@irin.univ-nantes.fr+

Abstract

Within the framework of Web information retrieval, this paper presents some methods to improve an indexing process which uses terminology oriented ontologies specific to a field of knowledge. Thus, techniques to enrich ontologies using specialization processes are proposed in order to manage pages which have to be indexed but which are currently rejected by the indexing process. This ontology specialization process is made supervised to offer to the expert of the domain a decision-making aid concerning its field of application. The proposed enrichment is based on some heuristics to manage the specialization of the ontology and which can be controlled using a graphic tool for validation.

Categories and Subject Descriptors: H.3.1 [**Content Analysis and Indexing**]

General Terms: Abstracting methods, Dictionaries, Indexing methods, Linguistic processing, Thesauruses

Additional Key Words and Phrases: Ontology, Enrichment, Supervised Learning, Thesaurus, Indexing Process, Information Retrieval in the Web

1 Introduction

Search engines, like Google[1] or Altavista[2] help us to find information on the Internet. These systems use a centralized database to index information and a simple keywords based requester to reach information. With such systems, the recall is often rather convenient. Conversely, the precision is weak. Indeed, these systems rarely take into account content of documents in order to index them. Two major approaches, for taking into account the semantic of document, exist. The first approach concerns annotation techniques based on the use of ontologies. They consist in manually annotating documents using ontologies. The annotations are then used to retrieve information from the documents. They are rather dedicated to request/answer system (KAON[3]...) The second approach, for taking into account of Web document content, are information retrieval techniques based on the use of domain ontologies [8]. They are usually dedicated for retrieving documents which concern a specific request. For this type of systems, the index structure of the web pages is given by the ontology structure. Thus, the document indexes belong to the concepts set of the ontology. An encountered problem is that many concepts extracted from document and which belong to the domain are not present in the domain ontology. Indeed, the domain coverage of the ontology may be too small.

In this paper, we first present the general indexing process based on the use of a domain ontology (section 2). Then, we present an analysis of experiment results which leads us to propose improvements of the indexing process which are based on ontology enrichment. They make it possible to increase the rate of indexed concepts (section 3). Finally, we present a visualisation tool which enables an expert to control the indexing process and the ontology enrichment.

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) [8]:

1. For each page, a flat index of terms 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 ([14]).

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 Ontology Cover Degree or OCD), the number of concepts included in the pages (the Direct Indexing Degree or DID and the Indirect Indexing Degree or IID)... The global evaluation of the indexing process (OSAD: Ontology-Site Adequacy Degree) is a linear combination of the previous measures (weighted means) among different threshold from 0 to 1. The measure enables us to quantify the “quality” of our indexing process (see [8]) for more details).

[1] http://www.google.com
[2] http://www.altavista.com
[3] http://kaon.semanticweb.org/

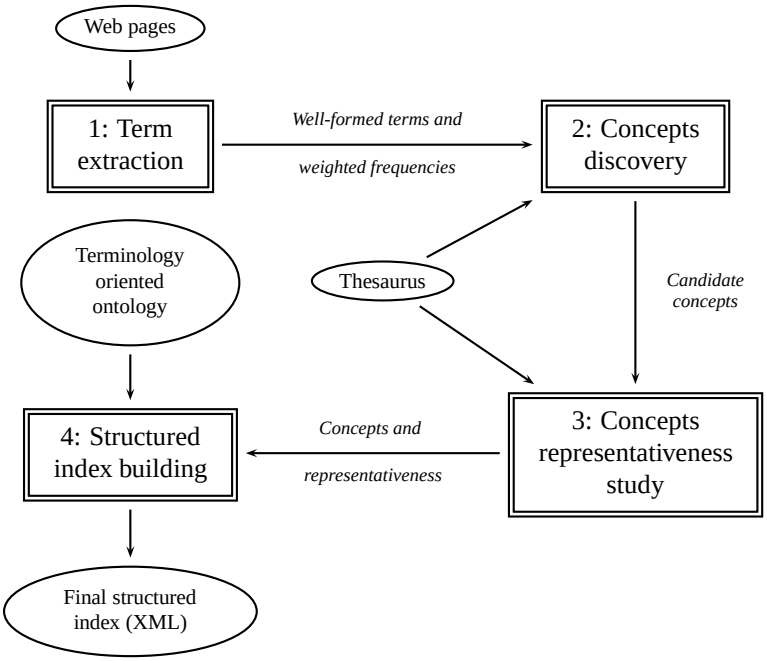

Figure 1: The indexing process

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 is 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);

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. 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...

Candidate concept status	Value
Valid and indexed (representativeness degree greater than 0.3)	**2 369**
Not indexed	337 428
Not in the ontology	333 547
With a representativeness degree greater than 0.3	**60 142**
Not in Wordnet	**4 049**
$< noun >$ of $< noun >$	293
$< noun >$	2 097
$< noun >< noun >$	1 414
$< noun >< noun >< noun >$	245
In Wordnet	**56 093**
With a representativeness degree lower than 0.3	273 405
Low representativeness degree (but in the ontology)	3 881
Number of processed candidate concepts	**339 797**

Table 1: Distribution of extracted concepts from 1000 Web pages of the site of the University of Washington (with a threshold of 0,3).

Our structured index allows us to build a new answering system using the "isa" relationship between concepts. Typically, when a request looks for a concept C, the system can reply not only pages containing C but also pages containing concept that are subsumed by C. Then, "not", "and" and "or" operators become operators over sets.

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 (using our different measures). This latter case enables us to classify a Web site in a hierarchical index like a classical search engine (Yahoo !, Excite...).

3 Techniques used to improve the general indexing process

The indexing process presented in the previous section takes into account the semantic of terms. Experiments carried out on various web sites, showed that, in this context, the information retrieval process precision is improved. However, in our approach, the final phase, which consists in matching candidate concepts with ontology concepts, is very basic. Indeed, only candidate concepts, belonging to the wordnet Thesaurus and whose representativeness degree is greater than a threshold, are taken into account. Among these concepts, only those belonging to the ontology are matched (4th phase of the general indexing process). In order to estimate the rate of concerned candidate concepts, we made analysis of our experiments results. We detail here these results carried out on the web site of the computer science department of the university of washington[4] (table 1).

339797 candidate concepts were processed. Only 2369 candidate concepts match with ontology concepts, while 337428 do not match. Among these rejected concepts, 3881 belong to the ontology but they are excluded because of their low representativeness degree. The 333 547 others are also rejected because they do not belong to the ontology. Among these candidate concepts not belonging to the ontology, 99,8% are rejected because of their low representativeness degree, while 0.18% of them, are only rejected because they do not belong to the ontology. The last rate seems to be low. However, if we compare the number of concepts that it represents (60 142) with the number of candidate concepts linked to the ontology concepts (2369), this rate becomes considerable.

We now study the set of these concepts. Among them, 93% belong to the Wordnet thesaurus but not to the ontology. This analysis lead us to improve the indexing process by enriching the ontology by candidate concepts belonging to the thesaurus and belonging to the domain covered by the ontology. The other part of concepts, 7% of them, belong neither to the ontology, nor to the wordnet thesaurus. But, some of them could be paired with ontology concepts. Some are acronyms of concepts which are related to concepts of the ontology. Others are noun

[4] "http://www.cs.washington.edu/" (1315 HTML pages).

phrases for which the head term is also matched with an ontology concept. Taking into account these two types of concepts will also improve the indexing process.

Works around ontology enrichment can allow us to improve our process [15, 19]. Two types of enrichments can be highlighted: the enrichment by refinement (or specialization) which tries to return a more specialized ontology and, the enrichment by abstraction [5] which tries to make the concerned ontology more general (by widening the field or by deleting too specific concepts). In our context (i.e. the use of a domain specific ontology), only the enrichment by specialization seems to be useful.

Consequently, we propose a semi-automatic method of ontology enrichment [12] that offers to the expert a powerful media to manage the covered domain by the ontology [17]. This management is made in three steps:

- an extension of our structured index building to add to the ontology concepts that seem to be "useful";
- a post-treatment based on a pruning of the final structured index (addition of a step 5 in the general indexing process, Fig. 2);
- a tool of ontology validation (section 4).

In the next section, we present this approach to improve the indexing process.

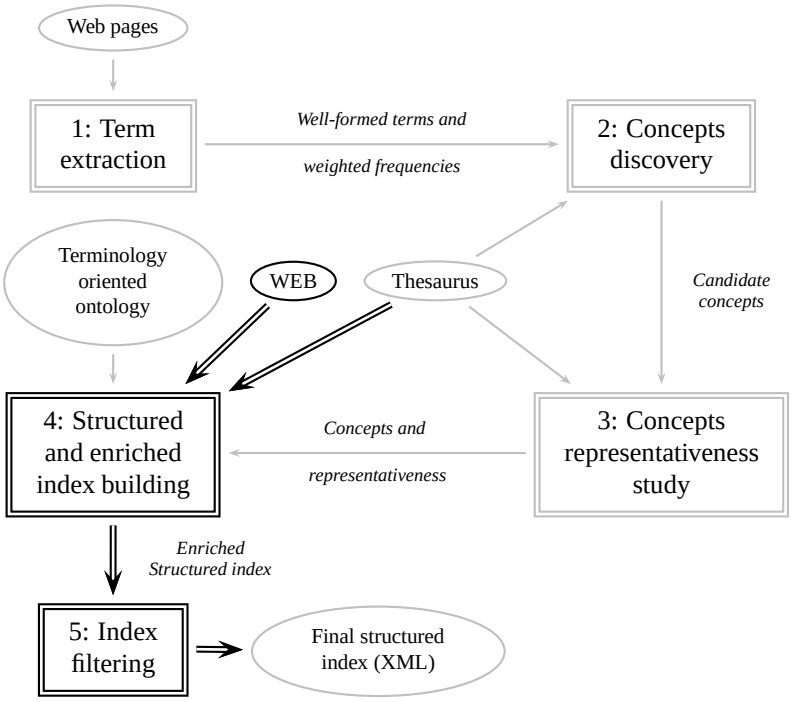

Figure 2: The "enriched" indexing process

3.1 The ontology enrichment using a thesaurus

The adopted method consists in exploiting a Thesaurus (Wordnet) during the building of the structured index (fig 1). This makes it possible to the process to add to the ontology some concepts which are present in the flat index, but which do not belong to the ontology. Indeed, we argue that, if a page of a Web site to be indexed relates quite to the field of the ontology used, a subset of the set of the concepts which represent its contents must be present in ontology. In order to determine the concepts which have to be added into the ontology, we use heuristics

based on hypernym paths associated to concepts in Wordnet and relative to ”isa” paths associated to concepts in the ontology.

These heuristics can de divided into two major categories (detailed later):

1. “usefulness heuristics”: i.e. only useful concepts are added like:
   (a) concepts that index at least one page (called *indexing concepts*);
   (b) concepts, which are useful to add properly indexing concepts (a concept which does not necessarily index a page but which subsumes one or more existing concepts and the new one). In other words, when a new indexing concept is added, new concepts can be added to update the ontology structure according to the structure of the thesaurus.
2. “specialization heuristics” (or “domain heuristics”):
   (a) new indexing concepts can cause a reorganization of the hierarchy according to the hypernym’s hierarchy of the thesaurus;
   (b) concepts are attached to the ontology only if they do not depend only on the top of the hierarchy. This rule enables the system to avoid the “rake” effect which implies an undesirable generalization of the ontology.

Now, we will detail some of these heuristics to show how our learning process takes into account the thesaurus hypernym hierarchy. The figure 3 presents the heuristics 1.a. Let us supposed that the concept c7#1, which is a Wordnet concept, is a concept belonging to our domain and its representativeness degree is greater than our threshold but it is not in our domain ontology. After the hypernym path generation via Wordnet (figure 3-(a)), the nearest concept which belongs to the ontology is c3#1 (figure 3-(b)). As this latter is not a root concept of our ontology (heuristics 2.b), c7#1 is attached to c3#1 without adding c8#2 and ci#j (figure 3-(c)) to the ontology.

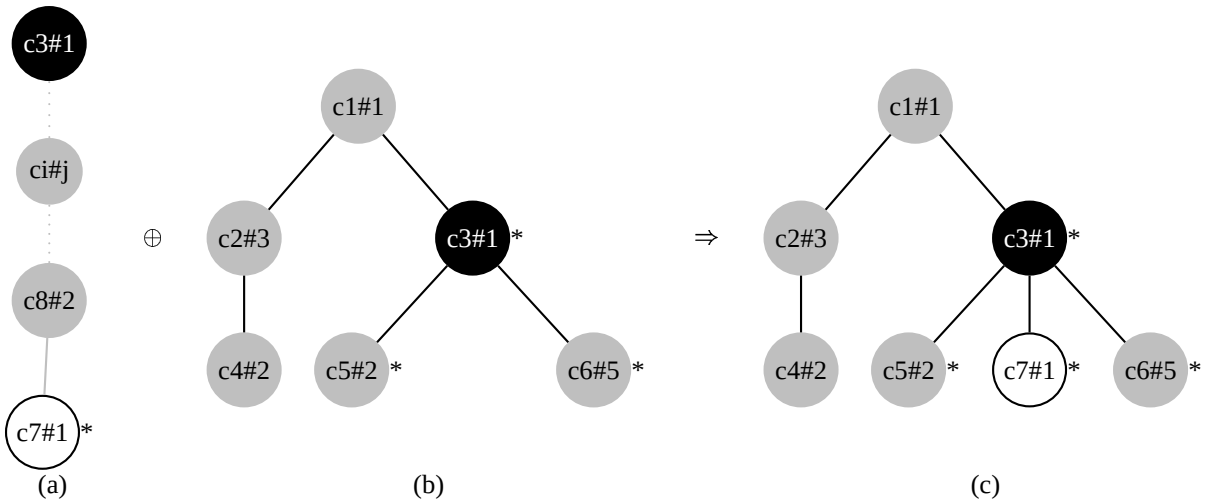

Figure 3: Adding a Wordnet concept to the domain ontology

Concerning the heuristics 2a, let us take our previous example. Let us suppose that we plan now to index concept c8#2 which subsumes the c7#1 one previously inserted. Then, an reorganization of the ontology structure is caused by this new concept insertion to be in agreement with the thesaurus hierarchy. Typically, c8#2 is attached to c3#1, and, consequently, concepts c7#1 and c6#5 are separated from c3#1 and are attached to c8#2. We suppose that this latter is an hypernym of them in our Thesaurus (fig.4).

Notice that the purpose of this set of rules is to guarantee a maximum indexing degree for this ontology (i.e. only the concept which indexes a page is added to the hierarchy).

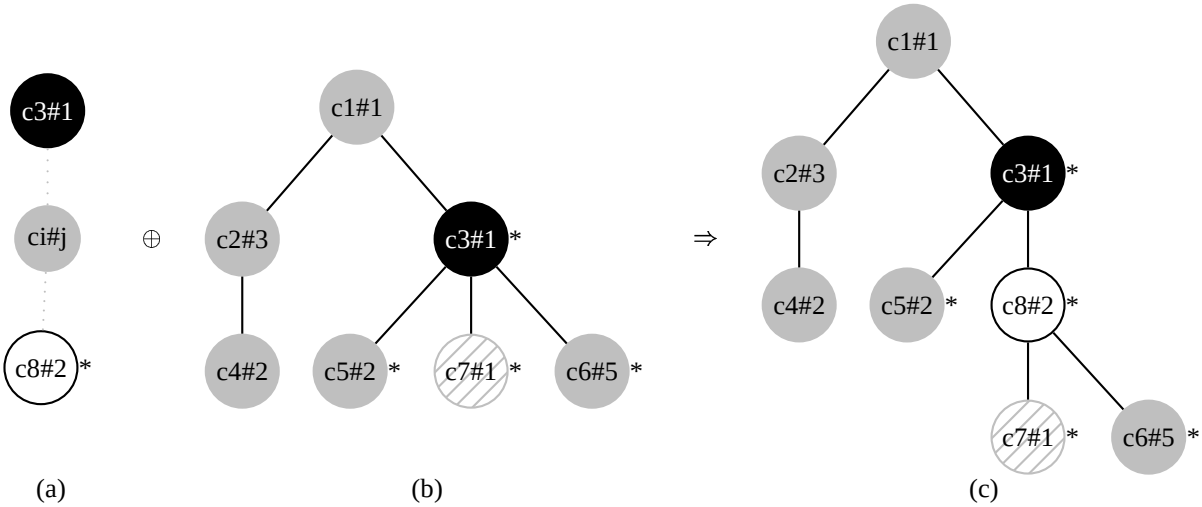

Figure 4: A process that adds a concept of Wordnet to the ontology with reorganization of the hierarchy.

3.2 The desambiguation of unknown terms

The method exposed in the previous sub-section, presents an automatic process to enrich the ontology of the domain by using a Thesaurus. So, new concepts are added in order to improve the indexing process, but only if they belong to the thesaurus.

Analyses of experiment results show that concepts not belonging to the thesaurus (but whose the representativeness degree is greater than the threshold) represent a great part of concepts that could be indexed (see Table 1). So this concept set decreases the indexing algorithm effectiveness. In order to take them into account, we have to develop other techniques for matching them (if possible) with ontology concepts.

Among these unknown concepts, we determine three categories of concepts:

1. Acronyms for which it is necessary to restore the initial form of the noun phrase;
2. Noun phrases whose the head term belongs to Wordnet.
3. Simple term which does not belong to Wordnet or noun phrases whose head term does not belong to Wordnet.

3.2.1 Acronym processing.

For acronyms processing (first category), we develop a technique to disambiguate them. It is based on the use of a Web acronym base[5]. This process restores their associated noun phrase. If the concept relative to the acronym belongs to the ontology, the acronym is added to the set of synonyms (synset) associated to this concept. In other case, if the concept associated to the acronym does not belong to the ontology, we can not attach it to the ontology and the term which represents it, is considered as an unknown term.

This method enables us to index terms like “cse” which often appear in our document. The “cse” term, for example, was found 392 times (there are 4049 unknown terms, see Table 1). It corresponds, in our context, to an acronym whose definition is: “Computer Science and Engineering”. Also, “uw” was met 165 times. It corresponds to “University of Washington”.

3.2.2 Noun phrases whose the head term is not in Wordnet.

The noun phrases associated to unknown concepts of the second category can be performed by syntactic analysis techniques like those proposed by [6] in order to determine their head term. If the head term already belongs to the

[5]like http://www.acronymfi nder.com, an online database that contains more than 277000 acronymes.

ontology, the noun phrase is inserted in ontology as its hyponym for example [18]. If the head term of the noun phrase belongs to Wordnet but not to the ontology, by the same process as that previously exposed in sub-section 3.1, we look for where to insert the concept relative to the head term in the ontology. And if it can be inserted, we add the noun phrase to the ontology as its hyponym for example. For example the noun phrase “department of computer science” is not in Wordnet, but “department” (the head term) is in the thesaurus. In this case, we add the noun phrase “department of computer science” to the ontology. It is inserted as an hyponym of the term “department”. Until now, our ontology only takes into account the “isa” relationship. Thus, we have not made a complete study of semantic relationships between noun phrase elements. Indeed, semantic relationships can be of various types (causality, composition...) [11].

3.2.3 Other unknown terms.

For unknown terms of the third category, namely simple term which do not belong to Wordnet or noun phrases whose head term does not belongs to Wordnet, we experiment the use of a method of refinement of ontology [4]. This method is based on topic signatures [1]. It enables to retrieve, via a search engine[6], a set of documents relative to a request containing a Wordnet synset which is an ontology concept [13]. The recovered document set is then analysed in order to extract, from these documents, terms which they contain and their occurence frequency. These terms are related to the Wordnet synset which is contained in the request. However, the name of the relationship which links them cannot be identified. This set of connected terms [10] will be used during the matching phase of the indexing process. If some of them are label of unknown concepts, they will be matched with the ontology concept contained in the request.

Obviously, and the first experiments prove it, a certain level of “noise” appears in collections. We plan, like [2, 3, 7], to implement rules in order to filter the signatures.

3.3 The management of ontology domain

The previous sub-sections have presented various techniques used to enrich the domain ontology during the Web site indexing process. These techniques make it possible to prevent a deviation of the application field of the ontology.

The deviation comes on the one hand from the lack of specialization of the Thesaurus, and on the other hand, it comes from the noise generated by the topic signatures techniques.

To this end, we propose a post-treatment technique based on the pruning of graph nodes which enables to control the ontology enrichment. This process proposes to human expert a set of concepts it can remove from the ontology. A visualisation tool which enables this task was developed (see Section 4).

To present the pruning rules which has been retained, we present an example of post-treatment (see Fig. 5).

Let us suppose that the C7#1 concept (Fig.5a) indexes 2 pages and C8#2 concept 7 pages. Both concepts are new ones. The pruning threshold (for instance 8 pages) indicates that if a concept indexes less than 8 pages, it will be pruned. Then, the C7#1 concept is not retained and the pages it indexes are attached to the C8#2 concept which subsumes it (Fig.5b). The C8#2 concept now indexes 9 pages (those coming from the C7#1 concept and those coming from the current indexing process). It is then kept and will be proposed to the human expert by the way of the OntologyManager visualisation tool (see section 4) for validation. Note that, even if the C6#5 concept and the C5#2 concept only index one page, they are kept because either they belong to the initial ontology or they was been added to the ontology and validated by a human expert.

This transfert of pages to a subsumer can be undo during a next indexing process which uses the same ontology. For instance, let us take the case presented in fig.4. But now, we make the assumption that c3#1 contains some pages (2 for example) coming from the concept c8#2. Then, after the addition of c8#2 concept in the ontology, these 2 pages go back to it (fig.6). So, if c8#2 is added with 6 pages attached to it in this new indexing process, this concept now indexes 8 pages (i.e. 6+2). Finally, it can be proposed to the expert and it will not be pruned.

[6] For instance, http://www.altavista.com, a search engine that allows keywords like AND, OR, NOT or NEAR.

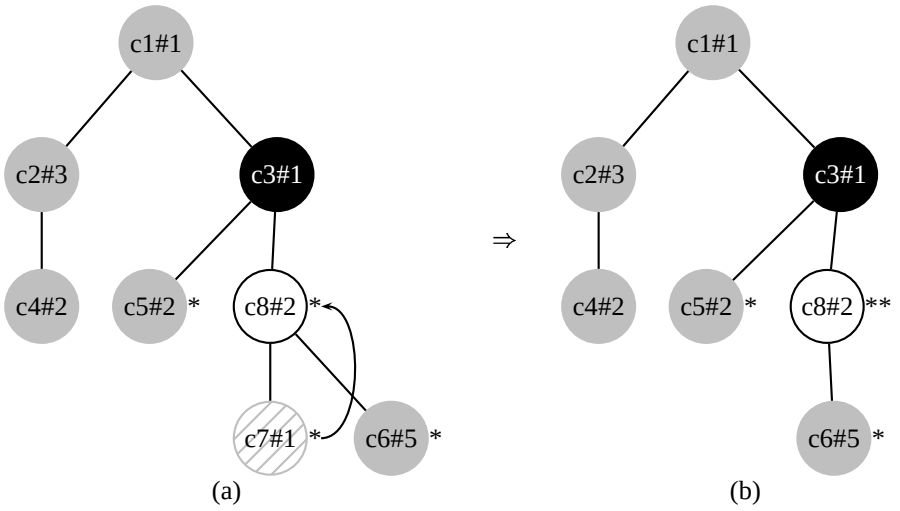

Figure 5: Post-treatment made on the final structured index.

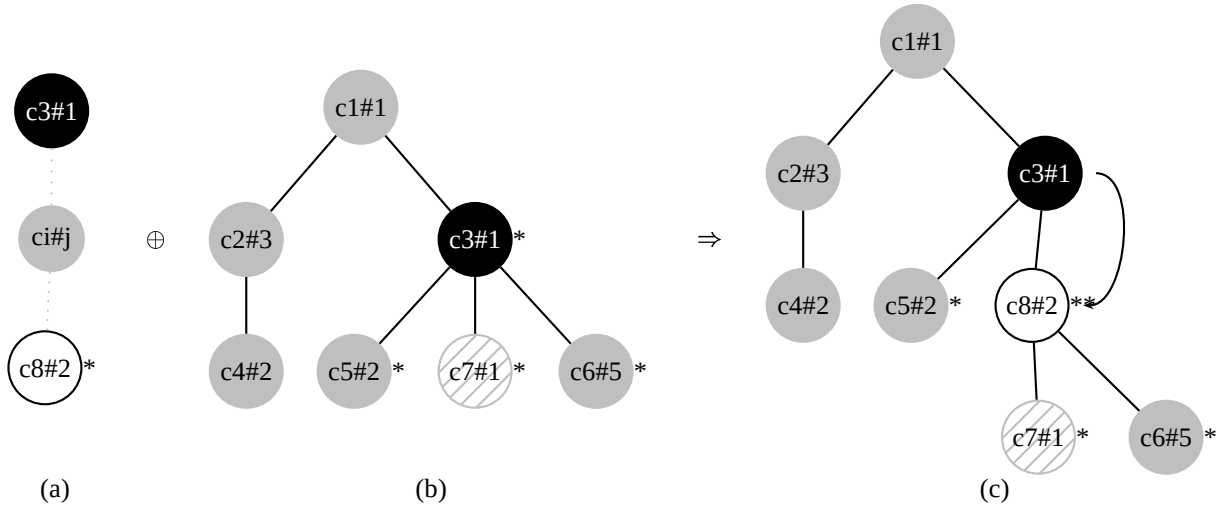

Figure 6: A process that adds a concept of Wordnet to the ontology with reorganization of the hierarchy and reassignement of the pages.

3.4 Results concerning the improvement of the indexing process

The table 2 shows results obtained after indexing of thousand pages of the site of the computer science department of the University of Washington. Results concerning three indexing methods are presented:

- the initial technique (only candidate concepts which match ontology concepts are retained);
- the ontology enrichment without pruning;
- the ontology enrichment with pruning.

We first notice, the great variation of the number of concepts added to the domain ontology between both methods of experiment which use enrichment. The method based on a simple enrichment enables to add to the ontology many concepts (80 concepts in the initial domain ontology, roughly 3500 concepts after the enrichment

Assessment	Initial indexing process	With the enrichment process	With the pruning process
Concepts belonging to initial ontology	80	3439	216
Coverage degree (OCD) (selected pages / pages to index)	84.33%	98.86%	98.86%
Indexing degree (DID) (selected concepts / concepts)	58.75%	99.04%	87.04%
OSAD Global evaluation	56.84%	67.62%	81.5%
Accepted candidate concepts	0.62%	11.15%	11.5%

Table 2: Results of the indexing process concerning 1000 pages of the site of the CSE department of the University of Washington (with a threshold of 0,3).

phases !). This phenomenon is due to the enrichment algorithm which authorizes the systematic addition of any representative concept (i.e. threshold of representativeness $\leq 0,3$) to the ontology of the domain. While the second enrichment method, which operates with pruning rules (see sub-section 3.3), enables to only add 136 concepts to the ontology.

Also let us notice that this method keeps the rate of coverage (98,86%) of the enrichment method without pruning. Indeed, during this pruning phase, some concepts which does not index enough pages (according to the threshold), are removed from the ontology. Their pages are then linked to concepts that subsume them.

Next, the number of concepts that index pages is growing. It is not surprising because we add only concepts indexing a minimal number of pages.

Finally, the rate of accepted concepts goes from 0.62% to 11.5% ! So, our process uses more available concepts that the pages contain.

4 OntologyManager: a user interface for ontology validation

A tool which makes it possible to control the ontology enrichment has been developed (see Figure 7). This tool implemented in java language, proposes a tree like view of the ontology. On the one hand, it proposes a general view of the ontology which enables the expert to easily navigate throw the ontology, on the other hand, it proposes a more detailed view which informs the expert about coefficient associated with concepts and pages. Notice that, in this last case, concepts are represented with different colours according to their associated coefficient. So a human expert easily can compares them. Moreover, some part of the ontology graph can also be masked in order to focus the expert attention on a specific part of the ontology. We are now developing a new functionality for the visualisation tool. It enables the user to have an hyperbolic view of the ontology graph (like OntoRama tool [9] or like H3Viewer [16]). In this context, the user can work with bigger ontologies.

The user interface also makes it possible to visualise the indexed pages (see Figure 8) and the ontology enrichment (by a colour system which can be customized). It will be easy to the human expert to validate or invalidate the added concepts, to obtain the indexing rate of a particular concept and to dynamically reorganize (by a drag and drop system) the ontology.

The concept validation process is divided into 4 steps defining 4 classes of concepts:

- bronze concepts: concepts proposed by our learning process and accepted by an expert just “to see”;
- silver concepts: concepts accepted by the expert for all indexing processes he/she does;
- gold concepts: concepts proposed by an expert to its community[7] for testing;

[7] A major definition of “ontology” is: a formal, explicit specification of a *shared* conceptualisation.

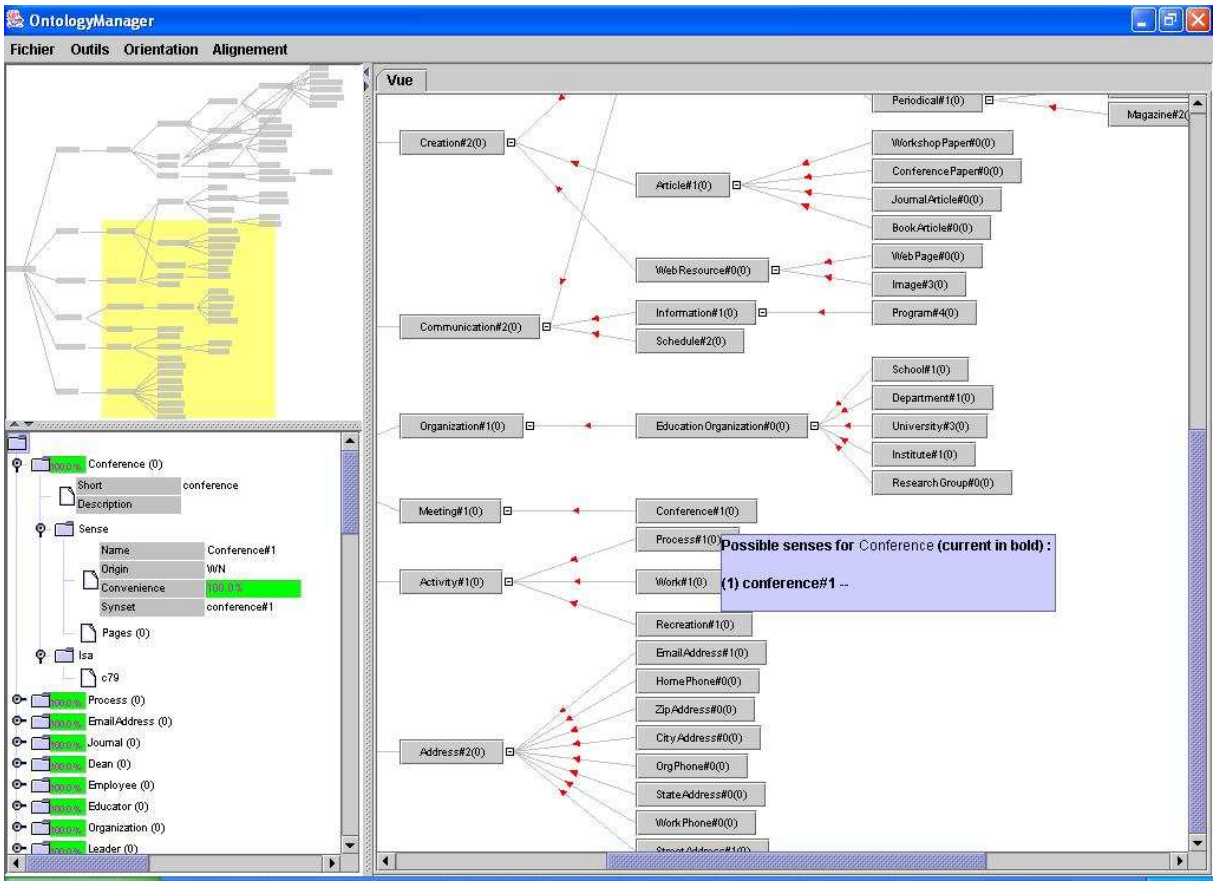

Figure 7: Visualisation of the ontology of the domain used for our experiments on the American Universities.

- platinium concepts: concepts accepted by the community.

All classes are managed by the expert by the way of the OntologyManager tool.

5 Conclusion

Available ontologies do not cover all the field of a domain. For example, some linguistic forms as acronyms, complex noun phrases are not belonging to the ontology. And these forms cause problems while the performing of an automatic content based indexing process. Indeed, these forms are not recognized by the system, thus they are not taken into account by the indexing process. In this paper, we present a method of ontology enrichment which makes it possible to take into account the concepts which do not belong to the ontology. It is based on the use of a thesaurus and of heuristics in order to add, to the ontology, concepts that belong to the thesaurus but not to the ontology. However, for terms neither belonging to the ontology nor to the thesaurus, we have developed a technique based on web requesting. The ontology enrichment is not fully automatic. A human expert will take the final decision to add or not a new concept to the ontology. To this end, we have developed a visualisation tool which help the human expert to control the indexing process and to control the potential ontology deviation.

Our future works are concerning the improvement of enrichment heuristics based on the addition to the ontology of concepts belonging to a thesaurus. Indeed, our heuristics are general. They would be more effective, if they were dedicated to particular set of concepts (depending of their depth in the ontology, their son number...). We are also working on the improvement of topic signature method. We also are making experiments to manage acronyms

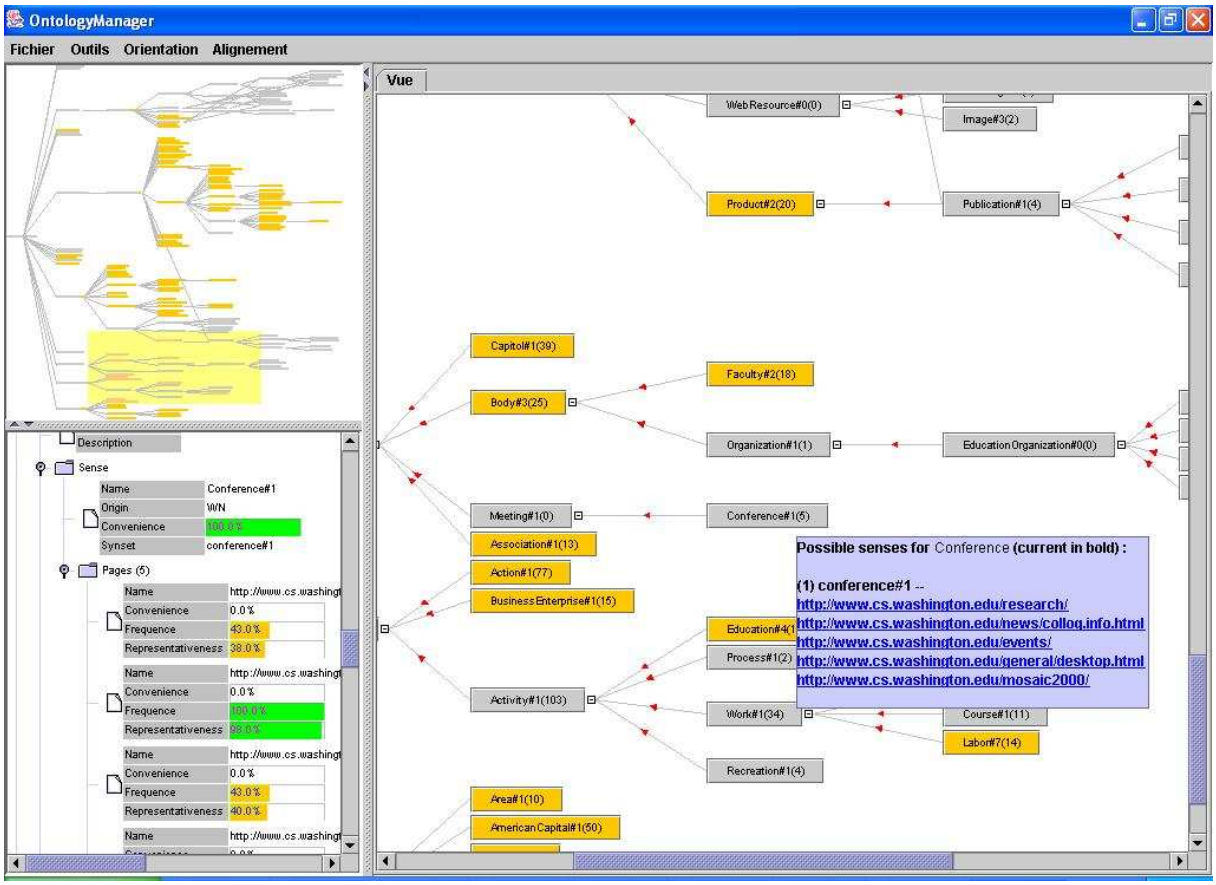

Figure 8: Visualisation of the final structured index concerning 1000 Web pages of the site of computer science department of the University of Washington.

not in the indexing phase but in the term extraction phase. For a given web pages, we are trying to determine to which noun phrase an acronym can be associated (which is the noun phrase which has its initial letters which correspond to the acronym). And, if the association cannot be made, we use a web acronyms base.

References

[1] AGIRRE, E., ANSA, O., HOVY, E., MARTINEZ, D.: Enriching very large ontologies using www. In Proceedings of the Ontology Learning Workshop ECAI, Berlin, Allemagne (2000)

[2] AGIRRE, E., ANSA, O., HOVY, E., MARTINEZ, D.: Enriching Wordnet concepts with topic signatures. In Proceedings of the SIGLEY Workshop on Wordnet and other Lexical Ressources: Applications, Extensions and Customizations, NAAL. (2001)

[3] ALFONSECA, E., MANANDHAR, S.: An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In Proceedings of the 1th International Conference on General Wordnet, Mysore, India.(2002)

[4] ALFONSECA, E., MANANDHAR, S.: Inproving an Ontology Refinement Method with Hyponymy Patterns. In Langage Resources and Evaluation, LREC-2002, (2002)

[5] ANTONIOU, G., KEHAGIAS, A.: On the Refinement of Ontologies. In Int. J. of Intelligence Systems. **15** (2000) 623–632.

[6] BOURIGAULT, D., JACQUEMIN, C., L'HOMME, M.-C.: Recent Advances in Computational Terminology. John Benjamins Publishing Company, Amsterdam/Philadephia, 2001

[7] BUITELAAR P. The SENSEVAL-2 Panel on Domains, Topics and Senses. In Proceedings of SENSEVAL-II. Toulouse, France, Aot.(2001)

[8] DESMONTILS, E., JACQUIN, C.: Indexing a Web Site with a Terminoly Oriented Ontology. IN: Cruz, I.F., Decker, S., Euzenat, J., McGuinness, D.L. (eds.): The Emerging Semantic Web. IOS Press, (2002) 181–197.

[9] EKLUND, P., ROBERTS, N., GREEN, S.: OntoRama: Browsing RDF Ontologies using a Hyperbolic-style Browser. In The First International Symposium on Cyber Worlds, CW02, Theory and Practices, IEEE Press. (2002) 405–411.

[10] FAATZ, A., STEINMETZ, R.: Ontology Enrichment with Texts from the www. In Semantic Web Mining, WS02, Helsinki, Finland. (2002)

[11] FABRE, C.: Interpretation of English nominal compounds: Designing a domain-independent model to guide semantic information retrieval. In Actes de CLIN-95 Anvers, Belgique.(1995)

[12] KIETZ, J., MAEDCHE, A., VOLZ, R.: A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In Aussenac-Gilles N., Biebow B., Szulman S., Workshop Ontologies and Texts.s, EKAW2000. Juan-les-Pins, France, 2-6 Octobre, (2000) 37–50.

[13] MIHALCEA, R., MOLDOVAN, D.I.: An Automatic Method for Genrating Sense Tagged Corpora. In Proceedings of the Conference of the American Association of Artificial Intelligence, AAAI99. Orlando, FL, juillet. (1999) 461–466.

[14] MILLER G. A.: Wordnet: an Online Lexical Database. In Int. J. of Lexicography, **3** (4), (1990) 235–312.

[15] MOTTA, E., BUCKINGHAM, SHUM S., DOMINGUE, J.: Ontology-Driven Document Enrichment: Principles and Case Studies. In 20th Workshop on Knowledge Acquisition, Modeling and Management, KAW99, Alberta, Canada, octobre. (1999)

[16] MUNZNER T. (1998). Drawing large graphs with H3Viewer and site Manager. In Proceedings of Graph Drawing, GW98. Montreal, Canada, August. (1998)

[17] OMELAYENKO, B. A.: Learning of Ontologies for the Web: the Analysis of Existent Approaches. In 8th International Conference on Database Theory, ICDT01. London, UK, 3 janvier. (2001)

[18] SEGUELA, P., AUSSENAC-GILLES, N.: Extraction de relations smantiques entre termes et enrichissement de modles du domaine. Actes de Confrences Ingnierie des Connaissances, IC?99. Palaiseau. (1999)

[19] SUMNER, T., DOMINGUE, J., ZDRAHAL, Z., HATALA, M., MILLICAN, A., MURRAY, J., HINKELMAN, K., BERNARDI, A., WESS, S., TRAPHNER, R.: Enriching Representations of Work to Support Organisational Learning. In Proceeedings of the Interdisciplinary Workshop on Building, Maintaining and Using Ogranizational Memories, 13th European Conference on Artificial Intelligence, ECAI98. 23-28 August, Brigthon, UK. (1998)

Ontology enrichment and indexing process

E. Desmontils, C. Jacquin, L. Simon

Abstract

Within the framework of Web information retrieval, this paper presents some methods to improve an indexing process which uses terminology oriented ontologies specific to a field of knowledge. Thus, techniques to enrich ontologies using specialization processes are proposed in order to manage pages which have to be indexed but which are currently rejected by the indexing process. This ontology specialization process is made supervised to offer to the expert of the domain a decision-making aid concerning its field of application. The proposed enrichment is based on some heuristics to manage the specialization of the ontology and which can be controlled using a graphic tool for validation.

Categories and Subject Descriptors: H.3.1 [**Content Analysis and Indexing**]

General Terms: Abstracting methods, Dictionaries, Indexing methods, Linguistic processing, Thesauruses

Additional Key Words and Phrases: Ontology, Enrichment, Supervised Learning, Thesaurus, Indexing Process, Information Retrieval in the Web

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