INFORMATION ACCESS VIA TOPIC HIERARCHIES AND THEMATIC ANNOTATIONS FROM DOCUMENT COLLECTIONS

Hermine Njike Fotzo, Patrick Gallinari
Université de Paris6 – LIP6, 8 rue du Capitaine Scott, 75015 Paris, France

Keywords: Concept hierarchies, typed hyperlinks generation, thematic annotations, text segmentation.

Abstract: With the development and the availability of large textual corpora, there is a need for enriching and organizing these corpora so as to make easier the research and navigation among the documents. The Semantic Web research focuses on augmenting ordinary Web pages with semantics. Indeed, wealth of information exists today in electronic form, they cannot be easily processed by computers due to lack of external semantics. Furthermore, the semantic addition is an help for user to locate, process information and compare documents contents. For now, Semantic Web research has been focused on the standardization, internal structuring of pages, and sharing of ontologies in a variety of domains. Concerning external structuring, hypertext and information retrieval communities propose to indicate relations between documents via hyperlinks or by organizing documents into concepts hierarchies, both being manually developed. We consider here the problem of automatically structuring and organizing corpora in a way that reflects semantic relations between documents. We propose an algorithm for automatically inferring concepts hierarchies from a corpus. We then show how this method may be used to create specialization/generalization links between documents leading to document hierarchies. As a byproduct, documents are annotated with keywords giving the main concepts present in the documents. We also introduce numerical criteria for measuring the relevance of the automatically generated hierarchies and describe some experiments performed on data from the LookSmart and New Scientist web sites.

1 INTRODUCTION

Large textual and multimedia databases are nowadays widely available but their exploitation is restricted by the lack of meta-information about their structure and semantics. Many such collections like those gathered by most search engines are loosely structured. Some have been manually structured, at the expense of an important effort. This is the case of hierarchies like those of internet portals (Yahoo, LookSmart, Infoseek, etc) or of large collections like MEDLINE: documents are gathered into topics, which are themselves organized into a hierarchy going from the most general to the most specific [G. Källgren, 1988]. Hypertext multimedia products are another example of structured collections: documents are usually grouped into different topics and subtopics with links between the different entities. Generally speaking, structuring collections makes easier navigating the collection, accessing information parts, maintaining and enriching the collection. Manual structuring relies on a large amount of qualified human resources and can be performed only in the context of large collaborative projects like e.g. in medical classification systems or for specific commercial products. In order to help this process it would be most needful to rely on automatic or semi-automatic tools for structuring document collections.

The Semantic Web whose goal is to help users to locate, organize, process information and compare documents contents, has for now focalised on the standardization, internal structuring of documents and sharing of ontologies in a variety of domains. The short-term goal is to transform existing sources (stored as HTML pages, in databases…) into a machine-understandable form. RDF (resources description form) has been created for computers but Semantic Web should be equally accessible by computers using specialized languages and interchange formats, and humans using natural language. Although the general framework of Semantic Web includes provisions for natural language technology, such techniques have largely been ignored. Nevertheless we can quote [B. Katz, J. Lin, 2002] who propose a method to augment

Njike Fotzo H. and Gallinari P. (2004).
INFORMATION ACCESS VIA TOPIC HIERARCHIES AND THEMATIC ANNOTATIONS FROM DOCUMENT COLLECTIONS.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 69-76
DOI: 10.5220/0002634000690076

RDF with natural language to be more familiar for users.

In this context, we study here how to automatically structure collections by deriving concept hierarchies from a document collection and how to automatically generate from that a document hierarchy. The concept hierarchy relies on the discovering of "specialization/generalization" relations between the concepts which appear in the documents of a corpus. Concepts are themselves automatically identified from the set of documents.

This method creates "specialization / generalization" links between documents and document parts. It can be considered as a technique for the automatic creation of specific typed links between information parts. Such typed links have been advocated by different authors as a mean for structuring and navigating collections. It also associates to each document a set of keyword representative of the main concepts in the document.

The proposed method is fully automatic and the hierarchies are directly extracted from the corpus, and could be used for any document collection. It could also serve as a basis for a manual organization.

The paper is organized as follows. In section 2 we introduce previous related work. In section 3, we describe our algorithm for the automatic generation of typed relations "specialization/generalization" between concepts and documents and the corresponding hierarchies. In section 4 we discuss how our algorithm answers some questions of the Web Semantic research. In section 5 we propose numerical criteria for measuring the relevance of our method. Section 6, describes experiments performed on small corpus extracted from Looksmart and New Scientists hierarchies.

2 PREVIOUS WORK

In this section we present related work on automatically structuring document collection. We discuss work on the generation of concept hierarchies and on the discovering of typed links between document parts. Since for identifying concepts, we perform document segmentation into homogeneous themes, we also briefly present this problematic and describe the segmentation method we use. We also give some pointers on work on natural language annotations for the Semantic Web.

Many authors agree about the importance of typed links in hypertext systems. Such links might prove useful for providing a navigation context or for improving research engines performances.

Some authors have developed links typologies. [Randall Trigg, 1983] proposes a set of useful types for scientific corpora, but many of the types can be adapted to other corpora. [C. Cleary, R. Bareiss, 1996] propose a set of types inspired by the conversational theory. These links are usually manually created.

[J. Allan, 1996] proposes an automatic method for inferring a few typed links (revision, abstract/expansion links). His philosophy is close to the one used in this paper, in that he chose to avoid complex text analysis techniques. He deduces the type of a link between two documents by analysing the similarity graph of their subparts (paragraphs). We too use similarity graphs (although of different nature) and corpus statistics to infer a relation between concepts and documents.

The generation of hierarchies is a classical problem in information retrieval. In most cases the hierarchies are manually built and only the classification of documents into the hierarchy is automatic. Clustering techniques have been used to create hierarchies automatically like in the Scatter/Gather algorithm [D. R. Cutting et al. 1992]. Using related ideas but by using a probabilistic formalism, [A. Vinokourov, M. Girolami, 2000], propose a model which allows to infer a hierarchical structure for unsupervised organization of documents collection. The techniques of hierarchical clustering were largely used to organize corpora and to help information retrieval. All these methods cluster documents according to their similarity. They cannot be used to produce topic hierarchies or to infer generalization/specialization relations.

Recently, it has been proposed to develop topic hierarchies similar to those found in e.g. Yahoo. As in Yahoo, each topic is identified by a single term. These term hierarchies are built from "specialization/generalization" relations between the terms, automatically discovered from the corpus. [Lawrie and Croft 2000, Sanderson and Croft 1999] propose to build term hierarchies based on the notion of subsumption between terms. Given a set of documents, some terms will frequently occur among the documents, while others will only occur in a few documents. Some of the frequently occurring terms provide a lot of information about topics within the documents. There are some terms that broadly define the topics, while others which co-occur with such a general term explain aspects of a topic. Subsumption attempts to harness the power of these words. A subsumption hierarchy reflects the topics covered within the documents, a parent term is more general than its child. The key idea of Croft and co-workers has been to use a very simple but efficient subsumption measure. Term *x* subsumes term *y* if the following relation holds :

$P(x|y) > t$ *and* $P(y|x)<P(x|y)$, where *t* is a preset threshold. Using related ideas, [K. Krishna, R.

Krishnapuram 2001] propose a framework for modelling asymmetric relations between data.

All these recent works associate the notion of concept to a term and rely on the construction of term hierarchies and the classification of documents within these hierarchies. Compared to that, we propose two original contributions. The first is the extension of these approaches to the construction of real concept hierarchy where concepts are identified by set of keywords and not only by a single term, all concepts being discovered from the corpus. These concepts better reflect the different themes and ideas which appear in documents, they allow for a richer description than single terms. The second contribution is the automatic construction of a hierarchical organization of documents also based on the "specialization/generalization" relation. This is described in section 3.

For identifying concepts, we perform document segmentation into homogeneous themes. We used the segmentation technique of [G. Salton et al. 1996] which relies on a similarity measure between successive passages in order to identify coherent segments. In [G. Salton et al. 1996], the segmentation method proceeds by decomposing texts into segments and themes. A segment is a bloc of text about one subject and a theme is a set of such segments. In this approach, the segmentation begins at the paragraph level. Then paragraphs are compared each other via a similarity measure.

For Now, Semantic Web Research focalises on the standardization, internal structuring of documents and sharing of ontologies in a variety of domains with the short-term to transform existing sources into a machine-understandable form (i.e. RDF). The researchers of the field realise that this language is not intuitive for common user and that it is difficult for them to understand formal ontologies and defined vocabularies. Therefore they preach as another mean of semantics augmentation, the annotation in natural language which is more intuitive for humans.

[B. Katz, J. Lin, 2002] propose a method to augment RDF with natural language to make RDF friendlier to humans and to facilitate the Web Semantic adoption by many users. Our approach is complementary to their work. Indeed, at the end of our structuring algorithm we have derived all the concepts present in the collection and for each document the set of concepts it is about. Then we are able to annotate documents with the set of its concepts. Each concept is represented by a set of keywords in the corpus language.

3 AUTOMATIC CONSTRUCTION OF TOPICS AND DOCUMENTS HIERARCHIES

3.1 Basic ideas

This work started while studying the automatic derivation of typed links "specialization / generalization" between the documents of a corpus. A link from document *D1* to document *D2* is of the type specialization (generalization from *D2* to *D1*), if *D2* is about specifics themes of *D1*. For example, *D1* is about war in general and *D2* is about the First World War in particular. This type of relation allows to build hierarchical organizations of the concepts present in the corpus which in turn allows for the construction of a hierarchical corpus organization.

In hierarchies like Yahoo!, the concepts used to organize documents are reduced to words. This gives only basic indications on the content of a document and the corresponding hierarchies are relatively poor. For this reason, we have tried to automatically construct hierarchies where each concept will be identified by a set of words. In order to do this, we need the knowledge of all themes present in the collection and of the specialization/generalization relations that do exist among them. From now on, we will identify a concept to a set of keywords.

For identifying the concepts present in a document, we use Salton segmentation method [G. Salton et al. 1996] which outputs a set of themes extracted from the corpus. Each theme is identified by a set of representative keywords.

For the detection of specialization/generalization relations between detected concepts, we will first build a term hierarchy like [Mark Sanderson, Bruce Croft, 1999], we then construct from that a concept hierarchy. After that, documents may be associated to relevant concepts in this hierarchy thus producing a document hierarchy based on the "specialization/generalization" relation between documents.

To summarize, the method is built around three main steps:

- Find the set of concepts of a given corpus
- Build a hierarchy (of type specialization /generalization) of these concepts
- Project the documents in the concepts hierarchy and infer typed links "specialization/generalization" between documents.

3.2 Algorithm

3.2.1 Concept extraction from a corpus

The goal here is to detect the set of concepts within the corpus and the words that represent them. For that, we extend Salton work on text segmentation:

We decompose a document into semantic themes using Salton's method [G. Salton et al. 1996], which can be viewed as a clustering on document paragraph.

Each document being decomposed in set of semantic themes, we then cluster all the themes in all documents to retain the minimal set of themes that ensure a correct coverage of the corpus.

We find for each concept the set of words that represent the concept. A concept is be represented here by its most frequent words.

3.2.2 Building the concepts hierarchy

The next step is to detect the "specialization/generalization" relations between extracted concepts so as to infer the concept hierarchy.

First, we build the hierarchy of terms within the corpus using Croft and Sanderson subsumption method [Mark Sanderson, Bruce Croft, 1999].

Then we create a concept hierarchy as follows. For each couple of concepts, we compute from the terms hierarchy the percentage *x* of words of concept *C2* generalized by words of concept *C1* and *y* the percentage of words of *C1* generalized by words of *C2*. If $x > S1 > S2 > y$ (*S1* and *S2* are thresholds) then we deduce a relation of specialization/generalization between these concepts (*C1* generalizes *C2*).[1]

After that, we have a hierarchical organization of concepts. It is therefore possible to attach indexed documents to the nodes in the hierarchy. One document may belong to different nodes if it is concerned with different concepts. Note that all concepts are not comparable by this "specialization / generalization" relation.

At this stage we already have an interesting organization of the corpus which rely on a richer semantic than those offered on classical portals or by term hierarchies [Lawrie and Croft 2000, Sanderson and Croft 1999]. However we can go further and establish "specialization/generalisation" links between corpus documents as explained below.

3.2.3 "Specialisation/generalization" relation between documents

Each document may be indexed by the set of corpus concepts and annotated by the set of keywords of the relevant concepts of its content. We then proceed in a similar way as for building concepts hierarchies from terms hierarchies:

For each couple of documents *D1*, *D2*, we compute from the concepts hierarchy the percentage of the concepts of D2 generalized by the concepts of *D1* and vice versa. This allows to infer a "specialization/generalization" relation between the two documents.

Note that it is a global relation between two documents, but we could also envisage relations between parts of documents. In particular, our "specialization/generalization" relation excludes the fact that two documents generalize one another which could happen when they deal with different concepts. *D1* could be a specialization of *D2* for concept *C1* and a generalization for concepts *C2*. However, we made this hypothesis for simplification.

Instead of building a hierarchy of documents, we could use the "specialization/generalization" relation to indicate links between documents. Such links could also be built between the document segments identified during the first step of the algorithm. This would result into an alternative representation of the document collection.

4 OUR ALGORITHM AND SEMANTIC WEB

This section will discuss how our algorithm answers some questions of Semantic Web research. The Semantic Web research can be view as an attempt to address the problem of information access by building programs that help users to locate, collect, and compare documents contents. In this point of view our structuring algorithm addresses some of these problems:

- The set of themes extracted from the collection, where each theme has a label in natural language, is a good synthesis of the collection for the user
- If one of the themes interests the user, he has all the documents treating the theme in the hierarchy node and each document has an annotation which reflects all the themes in it. This allows the user to target the subset of document likely to interest him.
- All documents are annotated by the themes they are about. The annotations are in natural

[1] Note that we briefly describe in section 6 an alternative method for directly building concept hierarchies without the need to first build the term hierarchy.

language and give a summary of the document.

Two others motivations are in the scope of Semantic Web research:

- Generate new knowledge from existing documents. Now this is not possible because computers cannot understand the contents of documents
- Synthesize and compile knowledge from multiples sources. To do this computers should be able to relate the contents of multiple documents, this is not the case.

For the second point typed links could be a part of the response. Indeed, typed links show a relation between documents contents and defined the relation. They are useful for users because they give them a navigation context when retrieving a corpus. We can also imagine that a search engine knowing the existing typed links could use them to improve the information retrieval.

In our algorithm we only derived the specialization/generalization link between documents contents. We are developing automatic method for other Trigg typology link typed.

5 EVALUATION MEASURES

Evaluating the relevance of a concept or document hierarchy is a challenging and open problem. Evaluations on user groups generally give ambiguous and partial results while automatic measures only provide some hints on the intrinsic value of the hierarchies. However, for avoiding at this stage the heavy process of human evaluation, we resort to automatic criteria to judge the quality of learned hierarchies. We therefore propose two measures of similarity between hierarchies. This will allow to compare the coherence of our automatic hierarchies to reference manual hierarchies (here a part of LookSmart hierarchy), but will not provide an indication of its absolute quality, neither will it tell us which hierarchy is the best.

5.1 A measure based on the inclusion

Documents in the hierarchy are said to share a relation of :

- "Brotherhood" if they belong to the same node
- "Parents-child" if they belong to nodes of the same branch

The first measure of similarity we propose is based on the mutual inclusion degree of hierarchies. The inclusion degree of hierarchy *A* with respect to hierarchy *B* is:

Inclusion(A,B) = $(N_f + Np)/(|F_A|+|P_A|)$

Where N_f is the number of couples of "brothers" in *A* which belong to *B*.

N_p is the number of couples "parents-child" in *A* which belong to *B*.

$|F_A|$ is the number of couples of "brothers" documents in *A*.

$|P_A|$ is the number of couples of "parents-child" in *A*

Finally, the similarity between *A* and *B* is the average of their mutual inclusion:

Similarity(A, B) = (inclusion(A, B) +

inclusion(B,A)) / 2

5.2 A measure based on Mutual Information

This similarity measure is inspired by the similarity measure between two clustering algorithms proposed in [T. Draier, P. Gallinari, 2001]. Let *X* and *Y* be the labels (classes) of all elements from a dataset according to the two different clustering algorithms and X_i be the label for the i^th cluster in *X*, $P_X(C = K)$ the probability that an object belongs to the cluster K in X, and $P_{XY}(C_X=k_x, C_Y=k_y)$ the joint probability that an object belongs to the cluster k_x in X and to the cluster k_y in Y. To measure the similarity of the two clustering methods, the authors propose to use the mutual information between the two probability distributions:

$MI(X,Y) = \Sigma_{i\in CX}\Sigma_{j\in CY} P_{XY}(C_X = i, C_Y = j)_* \log [(P_{XY}(C_X = i, C_Y = j)) / (P_X(C_X = i) * P_Y(C_Y = j))]$. If MI is normalized between 0 and 1 the more *MI(X, Y)* is close to *1* the more similar are the two set of clusters and therefore the methods.

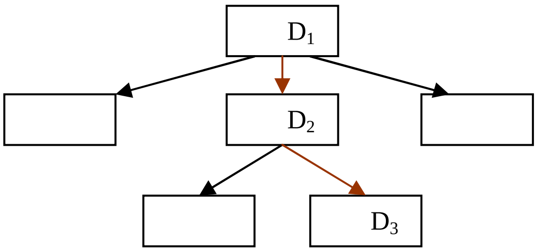

Figure 1: An example of documents hierarchy. We showed three nodes with only one document D_i. if we considered the node labelled D_3, it contains one document {D_3}, and for relation « parent-child » it contains the couples {(D_1, D_3), (D_2, D_3)}

In the case of hierarchical organization of documents, for measuring the similarity between two hierarchies, we need to measure how objects are grouped together (inside the hierarchy nodes) and to measure the similarity of the relations “parent-child” between objects in the two hierarchies. For simplifying the description, we will first consider that in each hierarchy one object may belong only to one node. The extension to the case where one object may appear in different nodes is easy but it is not exposed here.

For a hierarchy X let us note X_i a node of the hierarchy. A hierarchy of documents is described by two relations which are the relations “brotherhood” shared by the documents within a node and the relation of generalization between couples of documents sharing a relation of “parent-child”. A hierarchy can thus be seen like two simultaneous regroupings relating respectively on the documents and on the couples “parent-child”. The hierarchy is defined by the groups of documents which are linked by these two types of relation.

The mutual information *MI(X, Y)* between two hierarchies will be the combination of two components: $MI_D(X_D,Y_D)$ the mutual information between the groups of documents, corresponding to the nodes of the two hierarchies (it is the same measure as for a traditional clustering) and $MI_{P-C}(X_{P-C}, Y_{P-C})$ the mutual information measured on the groups of couples “parent-child” of the hierarchies. The mutual information between hierarchies X and Y will then be calculated by:

$MI(X,Y) = \alpha * MI_D(X_D,Y_D) + (1 - \alpha) * MI_{P-C}(X_{P-C}, Y_{P-C})$, *where* α is a parameter which allow to give more or less importance to the regrouping of documents in same the node or to the hierarchical relations “parent-child” documents. With this measure we can compare hierarchies of different structures.

6 EXPERIMENTS AND DATA

6.1 LookSmart and New-Scientist data

The data we used for our experiments are a part of the www.looksmart.com and www.newscientist.com sites hierarchies. First, we extracted a sub-hierarchy of LookSmart consisting of about 100 documents and 7000 terms about artificial intelligence. In a second experiment, we extract a sub-hierarchy of New-Scientist site consisting of about 700 documents. New-Scientist Web site is a weekly science and technology news magazine which contains all the latest science and technology news. Here the sub-hierarchy is heterogeneous sub-hierarchy whereas LookSmart data concern only AI. Documents are about AI, Bioterrorism, cloning, Dinosaurs, and Iraq. For each theme there are sub-categories concerning specifics aspects of the theme. In both cases, we compare the document hierarchies induced by our method and the term hierarchies to the original hierarchies, using the methods described in section 3.

6.2 Experiments and results

6.2.1 Segmentation, Annotations

In this section due to the place limitations we will give few examples of themes extract from the corpus, links between themes.

Comparing to the initial hierarchy of Looksmart with five categories, the hierarchy derived by our algorithm on the same corpus is more larger and deeper. Indeed, more of the original categories are specialized by our algorithm and it discovers new themes across the original ones. For example, many sub-categories emerge from “Knowledge Representation”: ontologies, building ontologies, KDD (where paper are about the data representation for KDD)… and most of the emerging categories are themselves specialized. In the same way, “Philosophy-Morality” is subdivided in many categories like AI definition, Method and stakes, risks and so on… Table 1 shows some examples of extracted themes on LookSmart data.

Table 1: examples of five concepts extracted from looksmart corpus, with a relation of generalization/ specialization between (2, 3), (2, 4), (2, 5)

1	Definition AI intelligence learn knowledge solve build models brain Turing test thinking machine
2	Informal formal ontology catalog types statement natural language name axiom definition logic
3	FCA technique pattern relational database data mining ontology lattice category
4	ontology Knowledge Representation John Sowa category artificial intelligence philosopher Charles Sanders Peirce Alfred North Whitehead pioneer symbolic logic
5	system KR ontology hierarchy category framework distinction lattice chart

Each document can then be annotated with the set of keywords of its index concepts (remember that after the step of concepts extraction all documents are indexed with their concepts).

6.2.2 Hierarchies similarities

In this section we compare the hierarchies induced by our method and term hierarchies to the original hierarchy using the measures of section 5.

Table 2: similarities between Looksmart and NewScientist data and others hierarchies (term, concept, concept_version2)

LookSmart			
	Terms.	Concepts1.	Concept2.
Inclusion	0.4	0.46	0.65
Mutual Information	0.3	0.6	0.7
NewScientist			
	Terms.	Concepts1.	Concept2.
Inclusion	0.3	0.2	0.6
Mutual Information	0.2	0.2	0.65

The concept hierarchy is large compared to the originals ones, and only a few documents are assigned to each concept. The greater width of the concepts hierarchy is due to the fact that some themes detected through corpus segmentation are not present in originals hierarchies which exploit poorer conceptual representations.

Nevertheless, the similarity between our hierarchy and LookSmart's is quite high. The inclusion similarity is about **0.5**, and the similarity based on the mutual information is around **0.6** (table 3). But the similarity between our hierarchy and New-Scientist one is low. This result point out the weakness of subsumption method our algorithm is based on, when the data are heterogeneous. We decide to modify our algorithm to be free from terms hierarchy induction for computing the subsumption relation between concepts. Remember that in our definition, concept *C1* subsumes concept *C2* if most terms of *C2* are subsumed by terms from *C1*. This relation was inferred from the term hierarchy (section 3.2.2). However it is also possible to directly derive the concept hierarchy without relying on the term hierarchy. For that we directly estimate P(concept C_i | concept C_j) by the number of documents containing both concepts divided by the number of documents containing concept C_j .

This hierarchy (denoted Concepts2. in table 2) seems closer to the manual hierarchy. It detects less subsumption relations between concepts on looksmart data; therefore it is less wide than the first concept hierarchy. Why does the number of subsumption between concepts fall down in the second method? A reason might be that in the construction of the first concept hierarchy, concept C2 is generalized by concept C1 if most of the terms of C2 are generalized by *C1* terms. Let us take the extreme case where only one word *w1* of *C1* generalizes the *C2* terms. In this case, we will say that *C1* generalizes *C2*. Actually, we can say that the presence of *C2* in a document implies the presence of *w1*, but it is not sure that it implies the presence of *C1*. For the second concept hierarchy the subsumption of *C2* by *C1* ensures that *C2* implies *C1*.

For the newscientist data, due to the heterogeneity of the vocabulary, subsomption test fail for many pairs of word and this effect is more drastic when projecting theme on term hierarchy. The consequence is that many theme nodes is compose by one document. Therefore the hierarchy is far from the original one. Modifying the definition of subsomption concept gives a hierarchy more similar than the original one. One way to reduce the influence of vocabulary heterogeneity is to consider synonyms in the computation of P(term1|term2).

These experiments shed some light on the algorithm behaviour. The hierarchies we obtain are coherent (particularly the second those obtain with the second method) compared to LookSmart and New-Scientists hierarchies, particularly on the groups of documents detected, but some of the documents pairs sharing the relation "Parent-Child" in the concept hierarchy do not appear in Looksmart hierarchy. This is inherent to the difference of nature between the two hierarchies.

If we compare the automatically built term hierarchy with that of LookSmart, we see that inclusion similarity is **0.4** and the mutual information is **0.3**. Both hierarchies use terms to index and organize documents. However, the term hierarchy uses all terms in the collection, whereas LookSmart uses a much smaller vocabulary. Therefore the hierarchy term is very large compared to LookSmart. Nevertheless some groups of documents are still common to the two hierarchies.

The similarity of the concept hierarchy with Looksmart seems higher than that of the term hierarchy.

7 CONCLUSIONS AND PERSPECTIVES

We have described a method to automatically generate a hierarchical structure from a documents collection. The same method can be used to build specialization/generalization links between documents or document parts and augmented documents with metadata in natural language. We have also introduced two new numerical measures for the open problem of the comparison and evaluation of such hierarchies. These measures give an indication on the proximity of two hierarchies; this allows measuring the coherence of two different hierarchies. On the other hand, they do not say anything on the intrinsic quality of the hierarchies. We are currently working on the development of measures for quantifying how much a hierarchy respects the "specialization/generalization" property.

Our method applied to LookSmart and New-Scientists data gives interesting first results although there is still place for improvements. The experiments also show that our concepts hierarchies are nearer to original hierarchies than a reference method which automatically builds terms hierarchies. Further experiments on different collections and on a larger scale are of course needed to confirm this fact.

We also show that our algorithm could give some answers to Semantic Web research concerns:

- Thematic hierarchies make easier information access and navigation, they are also a mean to synthesize a collection
- The algorithm allow to automatically related document sharing a specialization/generalization relation.
- At the end of the method each document is augmented with a set of keywords which reflects the concepts it is about

A perspective could be the use of automatically extracted concepts to build or enrich ontologies in a specific domain.

REFERENCES

J. Allan, 1996. Automatic hypertext link typing. Proceeding of the ACM Hypertext Conference, Washington, DC pp.42-52.

C. Cleary, R. Bareiss, 1996. Practical methods for automatically generating typed links. Hypertext '96, Washington DC USA

D. R. Cutting, D. R. Karger, J. O. Pedersen, J. W. Tukey, 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In ACM SIGIR.

T. Draier, P. Gallinari, 2001. Characterizing Sequences of User Actions for Access Logs Analysis, User Modelling, LNAI 2109.

G. Källgren, 1988. Automatic Abstracting on Content in text. Nordic Journal of Linguistics. pp. 89-110, vol. 11.

B. Katz, J. Lin, 2002. Annotating the Semantic Web Using Natural Language. In Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002) at COLING 2002.

B. Katz, J. Lin, D. Quan, 2002. Natural Language Annotations for the Semantic Web. In Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE2002).

K. Krishna, R. Krishnapuram, 2001. A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining. Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management. Atlanta, Georgia, USA. Pp.571-573

Dawn Lawrie and W. Bruce Croft, 2000. Discovering and Comparing Topic Hierarchies. In proceedings of RIAO 2000.

G. Salton, A. Singhal, C. Buckley, M. Mitra, 1996. Automatic Text Decomposition Using Text Segments and Text Themes. Hypertext 1996: 53-65

Mark Sanderson, Bruce Croft, 1999. Deriving concept hierarchies from text. In Proceedings ACM SIGIR Conference '99, 206-213.

Randall Trigg, 1983. A network-based approach to text handling for the online scientific community. University of Maryland, Department of Computer Science, Ph.D dissertation, November 1983.

A. Vinokourov, M. Girolami, 2000. A Probabilistic Hierarchical Clustering Method for Organizing Collections of Text Documents. Proceedings of the 15th International Conference on Pattern Recognition (ICPR'2000), Barcelona, Spain. IEEE computer press, vol.2 pp.182-185.