Learning « Generalization/Specialization » Relations between Concepts – Application for Automatically Building Thematic Document Hierarchies

Hermine Njike Fotzo, Patrick Gallinari
LIP6
8, rue du capitaine Scott
75015 Paris, France
{Hermine.Njike-Fotzo, Patrick.Gallinari}@lip6.fr

Abstract

We introduce a new method for automatically constructing concept hierarchies where the concept nodes follow a generalization / specialization relation. Starting from a set of concepts automatically extracted from a corpus, we show how to learn generalization / specialization relations between couples of concepts and how this leads to the construction of the hierarchy. We present an application of this method for building thematic document hierarchies similar in spirit to those found on internet portals. We also introduce new criteria for evaluating the quality of such hierarchies and for comparing them. We finally describe a series of tests performed on document collections coming from LookSmart and NewScientist hierarchies.

KEYWORDS
Structuring corpora, semantic relations between concepts, concept hierarchies, text segmentation

1. INTRODUCTION

Most available textual corpora are flat collections with no indication of semantic relations between documents or document elements. Some collections like those on internet portals (Yahoo, Open Directory, LookSmart…), are manually structured into thematic hierarchies by using extensive human resources. In these hierarchies, documents are gathered into topics, which are themselves organized into a hierarchy going from the most general to the most specific (Källgren, 1988) so that users can easily navigate in these hierarchies. Beyond navigation, structuring collections may help search engines, it makes easier accessing information parts and maintaining and enriching the collection.

We study here how to learn automatically concept hierarchies from documents collection according to a generalization/specialization relation. We propose a method which starts by automatically learning concepts from a corpus, and then learns generalization/specialization relations between these concepts. Such learned relations then allow creating a hierarchy of these ordered concepts.

We show as an example how to use this hierarchy of concepts for building a hierarchy of documents. The evaluation of the quality of such document hierarchies is an open problem and it is largely subjective. We introduce new criteria allowing the computation of quantitative indices measuring the quality of document hierarchies. We test the proposed methods on two corpora extracted from internet site hierarchies.

Our focus is on completely automatic methods for which there is no a priori available source of knowledge and everything has to be learned from the corpus. The method we propose is not aimed at inferring fine semantic knowledge at the sentence level like in terminology, nor do we rely on specific linguistic resource. On the opposite, our aim here is to build general tools able to extract simple semantic relations between corpus elements, which can be used in various information retrieval tasks.

The paper is organized as follows: in section 2 we describe related work. Section 3 introduces the main ideas behind the learning of "generalization/specialization" relations which relies on the subsumption relation described in (Croft & Sanderson, 99). Section 4 describes in detail the algorithm for extracting relations between concepts of a corpus. In section 5 we propose original numerical criteria to measure the quality of hierarchies. Finally section 6 describes experiments performed on the corpora extracted from two sites: LookSmart and NewScientist.

2. Previous Work

In this section we present work relevant to the generation of documents hierarchies and to the learning of specialization/ generalization relations between concepts. For automatically identifying concepts from corpora we perform the segmentation of documents into homogeneous themes, we also briefly present here the segmentation method we have been using.

2.1. Structuring of concepts and documents collection

Classifying documents within a hierarchy of concepts going from the most general to the most specific, has proved a reliable method for navigating a collection. In information retrieval several approaches were developed for the generation of hierarchies. In most cases, hierarchies are built manually and the classification of documents into the hierarchy is the only automatic step.

Clustering techniques have often been used to create documents hierarchies. These techniques use the similarity between documents generally measured using term frequency representations. Such hierarchies have been used to help navigation or retrieval. An example which is often quoted is Scatter/Gather (Cutting et al., 1992). This algorithm recursively clusters sets of documents in order to create a hierarchy. Using related ideas, (Vinokourov & Girolami, 2002] propose an unsupervised probabilistic model for inferring a hierarchical structure for the organization of document collections. Hierarchical clustering techniques have also been frequently used for organizing corpora. All these methods rely on the similarity between two documents or document sets. In this type of hierarchy, there is no semantic relation between the nodes at different levels. Also, they cannot be used to infer semantic relations between the concepts corresponding to the different clusters. As a consequence, they are practically useless for navigating collections.

Recently new types of hierarchies which are automatically built from corpora have been proposed (Lawrie & Croft, 2000; Lawrie et al., 2001; Sanderson & Croft 1999). They are term hierarchies built from generalization / specialization relations automatically discovered between terms in a corpus. Once this term hierarchy is built, it is possible "to project" documents on it, thus producing a document hierarchy. In the same spirit, (Krishna & Krishnapuram, 2001) propose a general framework for modeling asymmetrical relations between terms.

We propose to extend these approaches to the construction of real concept (themes) hierarchies 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. This is carried out by the automatic learning of "generalization/specialization" relations between these concepts which are themselves automatically discovered from the corpus. This allows in particular navigating a collection by visiting its different themes. From this hierarchy of concepts, one can generate "generalization/specialization" links between documents which can be an additional tool for navigating a collection.

There has been also a significant amount of work for the semi-automatic detection of semantic relations between terms in the area of terminology. To quote a few of them, let us mention the analysis of terms role in discourse, the construction of dictionaries and terminological reference resources for a given specific domain, the detection of semantic relations among terms (Morin, 1999; Morin & Jacquemin, 1999; Ruge, 1997). These relations operate at the sentence semantic level and are validated by human. Compared to this type of analysis, we seek for general relations between document parts, not between sentence terms, and develop a completely automatic process.

(Hernandez & Grau 2002; Hernandez & Grau, 2003] share many ideas with our approach. They make use of thematic text segmentation to highlight the text structure and describe the text segments by identifying their topics and their roles. Topic description is performed at a global and a local level. Thematic segments are represented by nominal groups, the importance of which for representing a theme is computed locally with respect to all thematic segments and globally with respect to the whole document. This structuring facilitates intra-documents navigation and the automatic formatting of electronic documents for a targeted visualization. In contrast with this approach, our aim is the structuring of whole collections and not of individual documents.

2.2. Text Segmentation

The segmentation task consists in identifying homogeneous text regions or frontiers corresponding to topic shifts between such regions. There is a large body of literature in this area since this could be used for many different applications in information retrieval (Kalvans et al., 1998). There is a variety of related but different segmentation problems depending on the task it is used for. We consider here the segmentation of documents into topically coherent and homogeneous passages (Hearst, 1997) for identifying themes in a collection. We used for that the thematic segmentation technique proposed by (Salton et al., 1996). This method proceeds by decomposing texts into segments and themes, a segment being a bloc of contiguous text about one subject and a theme a set of such segments. It makes use of classical similarity measures between segments. The hypothesis is that if the representations of two text segments have a weak similarity then these segments should have few thematic links. Segmentation starts at the paragraph level which is justified by the fact that authors generally expose one point of view per paragraph. It then proceeds as follows:
- Compute the similarities between paragraphs in a document and retain those higher than a given threshold. Construct the graph of similarities and extract triangles from this graph. A triangle is a set of three paragraphs with strong similarities and therefore susceptible to represent a coherent topic.
- For each triangle, build its vector representation which is the average of the three vectors representing the paragraphs of the triangle.
- Merge the triangles whose similarity is higher than a given threshold. Repeat until convergence.

3. "Generalization / Specialization" Relation

In this section, we introduce the method described in (Croft & Sanderson, 99) for inferring generalization/ specialization relations between terms and for deriving term and document hierarchies. In section 4 we generalize this method to the inference of generalization/ specialization relations between more sophisticated concepts identified by a set of keywords instead of a single word.

There exists a "generalization/specialization" relation between entities D1 and D2, (for example D2 is a specialization of D1 and D1 a generalization of D2) if D2 evokes a specificity of D1, or is about specifics themes of D1. For example D1 = sport and D2 = football, or D1 is about the war in general and D2 treats First World War. This type of relation makes possible to build a hierarchical

organization of concepts present in a corpus and to derive from that a hierarchical organization of the collection documents.

Other types of relations can lead to a hierarchical organization of concepts and documents (for example the "pre-necessary" relation), but the "generalization/specialization" relation or "is –a" relation is very intuitive for users and is the most often used for structuring collections. It is used for example for organizing documents on internet portals such as Yahoo, Open Directory, LookSmart… and is also one of the major relations used in ontologies together with co-hyponymy, synonymy, antonymy…

3.1. Detecting "generalization/specialization" relation between two concepts

Most document hierarchies make use of simple concepts which could be identified by a single word. In general, the hierarchy of concepts (words) is built manually and only the classification of documents in this hierarchy is automatic. Recently (Croft & Sanderson, 1999) proposed a method for inferring term hierarchies automatically by learning a generality/specificity relation between terms; it is based on term subsumption. The idea is that some of the terms which occur frequently in a collection give significant information about the topics treated in the corpus. These terms may define a subject in a general way, whereas others which co-occur with these general terms and are less frequent explain some aspects of the subject. Subsumption tries to highlight the characteristics of the concepts and their relations.

The key idea of Croft and co-workers has been to use a simple but efficient subsumption measure, it characterizes a relation of generality/specificity between two terms and is based on asymmetrical term co-occurrence:

Term x subsumes term y (or x is more general than y) if the following relation holds $\Leftrightarrow P(x/y) > t$ and $P(y/x) < P(x/y)$, where t is a preset threshold.

$n(x,y) =$ is the number of documents that contain terms x and y.

$n(y) =$ is the number of documents that contain term y.

$P(x/y) = n(x,y) / n(y).$

In others words, *x* subsumes *y* if documents in which *y* occurs are a subset or nearly a subset of the documents in which *x* occurs. The second rule ensures that if both terms occur together more than *t%* of the time, the most frequent term will be chosen as the general concept.

This term subsumption measure can be extended to topic subsumption, where each topic is represented by a set of keywords as we will see in section 4.

3.2. Remarks on the subsumption measure and similar statistics

The first important remark concerning this subsumption measure is that it is suitable in domains where terms are often repeated. If this was not the case, the co-occurrence estimations would not be robust enough to be relevant. This definition will be valid for example for scientific corpora which make use of a reduced vocabulary and much less for literary corpora or newspapers articles. By using linguistic resources like WordNet to take into account synonymy, one can reduce the sensitivity of this technique to the variability of the corpora. However, this idea is better suited to homogeneous corpora where all the documents cover the same subject than to heterogeneous corpora. With this definition, a concept

can have several parents, this corresponds to the different senses of this concept and reflect its polysemia.

In this paper we extend this "generalization/specialization" relation to whole topics identified by sets of representative words. Others types of relations can be extracted by exploiting similar statistics, but require additional information sources to name them. For example, collocations within a sentence often model symmetrical relations (synonym, co-hyponym…). In the same way next-neighbour collocations model anti-symmetrical relations like hyponymy, class-instance… All these instances of collocation give the intuition of a relation between two entities, but do not allow naming it without additional information or analysis.

4. Extracting Relations and Generating the Hierarchies

Concepts which can be identified by a single word are simple ones, and lead to hierarchies which are rather crude. Furthermore, automatic methods like the one by Croft and Sanderson heavily rely on word repetition and co-occurrence, and fails on heterogeneous collections or when the vocabulary is important.

In this section, we show how to extend the ideas introduced in section 3 to the discovering of "generalization/specialization" between concepts identified by sets of representative words. We then describe how to build hierarchies of such concepts. Tests described in section 6 show that the corresponding hierarchies are much richer, they allow to capture more complex relations between concepts and are much less sensitive to the vocabulary and heterogeneity of corpora. We first present how concepts are identified and describe after that two methods for identifying generalization/ specialization relations between such concepts. The first one is an extension of Croft and Sanderson technique which characterizes theme subsumption from their representative terms subsumption. The second one directly identifies theme relations without using term subsumption. As will be seen in the experiments described in section 6, the latter method allows to infer high quality hierarchies.

4.1. Pre-processing and documents representation

Collections are first pre-processed via stemming and elimination of rare words. Let $V = \{w_j\}_{j \in \{1, \ldots, M\}}$ denote the vocabulary, $D = \{D_i\}_{i \in \{1, \ldots, N\}}$ the set of documents in the collection, $P = \{P_k\}_{k \in \{1, \ldots, L\}}$ the set of document paragraphs. Paragraphs will be the basic text unit for the segmentation step. Document D_i will be classically represented by a vector of frequencies:

$Di = (tf_i(w_1)*idf(w_1), \ldots, tf_i(w_M)*idf(w_M))$,
where $tf_i(w_j)$ is the frequency of term j in D_i, $idf(w_j) = log(N/df(w_j))$, with N the number of documents in the collection and $df(w_j)$ the number of documents containing w_j.

In the same way paragraph P_k is represented as:
$P_k = (tf_k(w_1)*ipf(w_1), \ldots, tf_k(w_M)*ipf(w_M))$,
where $tf_k(w_j)$ is the frequency of term j in P_k. $ipf(w_j) = log(L/dp(w_j))$, with L the number of paragraphs in the collection and $pf(w_j)$ the number of paragraphs containing w_j.

The similarity measure used between two entities (documents or paragraphs) is the cosine between the vectors of their characteristics.

4.2. Extracting concepts from the corpus

For extracting a set of concepts from the corpus and their set of representative words, we extend the segmentation method of (Salton et al., 1996): initially, we decompose each document into a set of semantic topics using Salton's method. We then cluster these topics in order to identify a representative set of topics for the corpus:

- We build a graph of similarity between the topics identified using Salton method on each document (there is an edge between two topics if their similarity is higher than a given threshold).
- We then compute the connected components of this graph. For each component, we keep only the nodes which are connected to at least 75% of the others nodes of the component.
- A component with at least β% of its documents (β is a threshold which has been fixed around 90% in our experiments) in a second component will be merged with the latter.
- The different resulting components form the set of corpus themes.

Each theme is represented by a set of keywords. In our experiments, we have used the most frequent words in the theme. From now on we will identify the "concepts" to these sets of keywords.

4.3. Inferring « generalization/specialization » relations between concepts

We now introduce two methods for computing generalization/specialization relations between concepts and to construct concept hierarchies.

4.3.1. Method 1: exploitation of Croft et al. terms hierarchy

The first method we propose detect relations between concepts by exploiting the terms hierarchy of (Croft & Sanderson, 1999). The concepts hierarchy is built as follows: For each couple of concepts (C1, C2), we compute from the term 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)*

This method inherits the weaknesses of Croft et al. method. In particular, it works only on homogeneous corpora with an important term repetition. In order to correct these weaknesses, we propose a second method.

4.3.2. Method 2: direct application of subsumption definition to concepts

The second approach consists in computing directly the conditional probabilities *P(Ci/Cj)* without using the word conditional probabilities (word subsumption). Estimating these probabilities for any pair of concepts allows applying the subsumption definition directly to the concepts. Once the relations of "generalization/specialization" are detected on the couples of concepts, we apply transitivity to build the concept hierarchy. Using this hierarchy, we can index the documents by the topics they contain and assign them to different nodes. A document can belong to several nodes if it treats several topics.

P(Ci/Cj) can be estimated by counting:
$P(C_i/C_j) = ($ *number of documents about concepts C_i and C_j)/(number of documents about C_j)*

For deciding whether a document D is about concept C or not, one needs to estimate $P(C/d)$ the probability of concept C for d which is not trivial.

The results of the document segmentation can be used to assign the concepts to the documents. If a paragraph in d belongs to concept C then $P(C/d)$ is non zero. It can be set to a real value, by measuring e.g. the importance of this paragraph in the document. This provides a crude estimation of $P(C/d)$ and many documents dealing with the concept but which does not have a whole paragraph associated with this concept will be ignored.

We rather propose to proceed to the estimation of $P(C/d)$ via an Estimation Maximization (EM) algorithm which is described below. This algorithm iteratively finds the $P(t/C)$ for all concepts C and vocabulary term t, by maximizing the likelihood of the document collection. Assuming a naïve Bayes model for the documents, this allows to compute $P(d/C)$ and therefore $P(C/d)$ via Bayes rule.

EM Algorithm:

Parameters : $P_t^C = P(t \in d \mid d \in C) = P(t \mid C)$

                         $P(t)$ = # docs containing t / # docs

                         $P(C/d)$

Initializations :   initialize P_t^C with the knowledge of concepts keywords

                         P_t^C = # of terms t in concept C / # of terms in concept C

Step E:

                         $P(C \mid d) = [P(C)/P(d)] * \prod_{t \in d} [P(t \mid C)]$

Step M: re-estimation of P_t^C with the results of step E

                         P_t^C = # of documents $\in C$ with term t / # documents $\in C$

                         For which we use $d \in C \Leftrightarrow P(C \mid d) >$ threshold

       The log likelihood maximized by this algorithm is

      $Log(L) = Log(P(D \mid \Theta)) = \sum_d \sum_{t \in d} log(\sum_C P(t \mid C, \Theta) P(C \mid \Theta))$

      where D is the corpus and Θ the model parameters.

5. Evaluation Measures

Evaluating the relevance and the quality of a hierarchy is a challenging task and remains for now an open problem. Evaluation by humans is a lengthy and costly process whose results may be difficult to interpret, ambiguous and subjectives. As for most IR tasks, quantitative criteria for automatic evaluation only provide partial information on the quality of the hierarchy. Nevertheless, they provide useful hints for comparing and evaluating learned theme hierarchies and quantifying the relevance of the relations discovered between concepts.

We propose below two original and complementary measures for evaluating theme hierarchies. The first is an indicator of similarity between hierarchies. This will enable us to compare the coherence of our automatic hierarchies with respect to existing manual hierarchies. This is not an indicator of the intrinsic quality of a hierarchy. Our second measure reflects how a hierarchy respects the generalization/specialization relation between its nodes.

5.1. Computing a similarity measure between documents hierarchies

Documents in a hierarchy are said to share a relation of "brotherhood" if they belong to the same node or a "parent-child" relation if they belong to nodes on the same branch. The similarity measure we propose is based on mutual information between hierarchies and is inspired by the similarity measure between cluster sets proposed in (Draier & 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 ith cluster in X, $P_X(C = K)$ the probability that a document belongs to cluster K in X, and $P_{XY}(C_X=k_x, C_Y=k_y)$ the joint probability that a document belongs to cluster k_x in X and to cluster k_y in Y. In order to measure the similarity of the two clustering methods, the authors propose to use the mutual information between the two probability distributions P_X and P_Y:

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

For comparing two document hierarchies, we need to compare simultaneously how documents are grouped together inside the hierarchy nodes and how similar are the "parent-child" relations between documents in the two hierarchies. For simplifying the description, we will first consider that in each clustering one document may belong only to one cluster. The extension to the case where one object may appear in different nodes is easy but it is not detailed here.
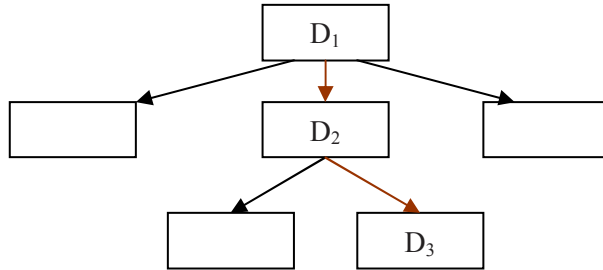

Figure 1: An example of document hierarchy. We showed three nodes with only one document D_i. if we considered the node labelled D_3, it contains one document $\{D_3\}$, and for relation « parent-child » it contains the couples $\{(D_1, D_3), (D_2, D_3)\}$.

Let X_i denote a node of hierarchy X. 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 as two simultaneous clustering operating 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 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 allows to give more or less importance to documents in the same node or to the hierarchical relations "parent-child".

This measure allows comparing hierarchies with different structures. In particular, the relative contribution of the two terms in $MI(X,Y)$ gives information about the similarity of document nodes and of "parent-child" relations between documents.

The measure does not take into account the depth in the generalization relation. Moreover brotherhood relations are considered only for pair of nodes and not in a more global way. Two hierarchies can be evaluated similar according to this measure when they present very different characteristics. For example, we can split a node of a given hierarchy into a series of nodes corresponding to the document pairs it contains and the similarity will be 1 whereas, the first hierarchy is more synthetic and probably more useful than the second one.

5.2. Quantification of the « specialization/generalization » capacity of a hierarchy

The second measure we propose quantifies how a hierarchy respects the relation of generalization/specialization between the objects which are in the nodes. It is based on the conditional entropy. Conditional entropy measures the uncertainty on a variable given the knowledge of another variable: $H(Y/X) = -\Sigma_x \Sigma_y P(x, y) * log(P(y/x))$.

Within the framework of subsumption, if a term x generalizes a term y then the uncertainty on x knowing y is low. Let us consider term hierarchies, we note GIT the Generalization Index for a Term and GIH the Generalization Index for a Hierarchy. They are defined as follows:

- $GIT(t, \{s_i\}) = \Sigma s_i - P(t, s_i) * log(P(t/s_i))$, where t is a term and the s_i are its sons. The lower this index, the better t generalizes its sons $\{s_i\}$.
- $GIH = \Sigma_{node} IGT(node, \{sons\}_{node})$.

This measure can be trivially extended to theme hierarchies.

6. Experiments and Results

6.1. Data

The data we used for our experiments are a part of the www.looksmart.com and www.newscientist.com site hierarchies. The first dataset consist of about 100 documents and 7000 terms about artificial intelligence (see figure 2) and is a homogeneous set of documents. The second one consists of about 700 documents and 20000 terms. The latter is a weekly science and technology news magazine which contains the latest science and technology news. The site is organized in hierarchies of themes. We extracted a heterogeneous sub-hierarchy from this site with documents about very different topics like 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 hierarchies induced by our methods to the original hierarchies on the same data by using the measures described in section 5.
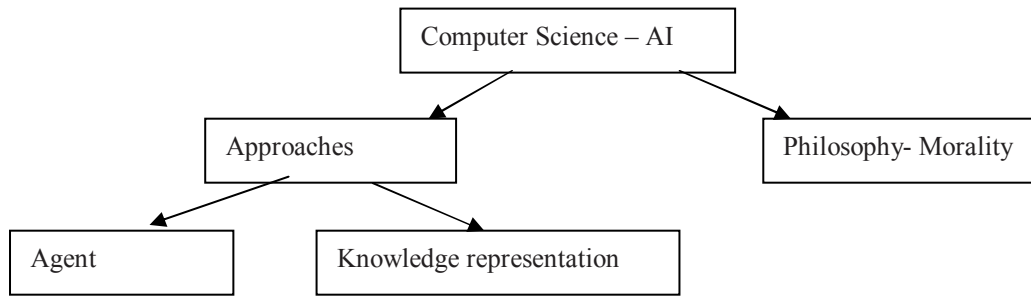
Figure2: Example of hierarchy – Sub-hierarchy of LookSmart site used in experiments

6.2. Experiment results

6.2.1. Example of extracted concepts and relations

In table 1, we give examples of extracted concepts on the LookSmart corpus. Each concept is identified by a set of keywords. Using the algorithm which directly infers $P(C_i|C_j)$ (section 4.3.2) relations of "generalization/specialization" between the concepts (2,3), (2,4), (2,5) are discovered.

Compared to the initial Looksmart hierarchy with five categories, the hierarchy derived by our algorithm on the same corpus is much larger and deeper. Most of the original categories are refined by our algorithm. For example, many sub-categories do emerge from the original "Knowledge Representation" category: ontologies, building ontologies, KDD (papers about 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…

LookSmart	
1	definition AI intelligence learn knowledge solve build models brain Turing Test thinking machine
2	informal formal ontology catalog types statements natural language names axiom definition logic
3	FCA techniques pattern relational database data mining ontology lattice categorie
4	ontology Knowledge Representation John Sowa categories artificial intelligence philosophers Charles Sanders Peirce Alfred North Whitehead pioneers symbolic logic
5	system KR ontology hierarchy categories framework distinction lattice chart

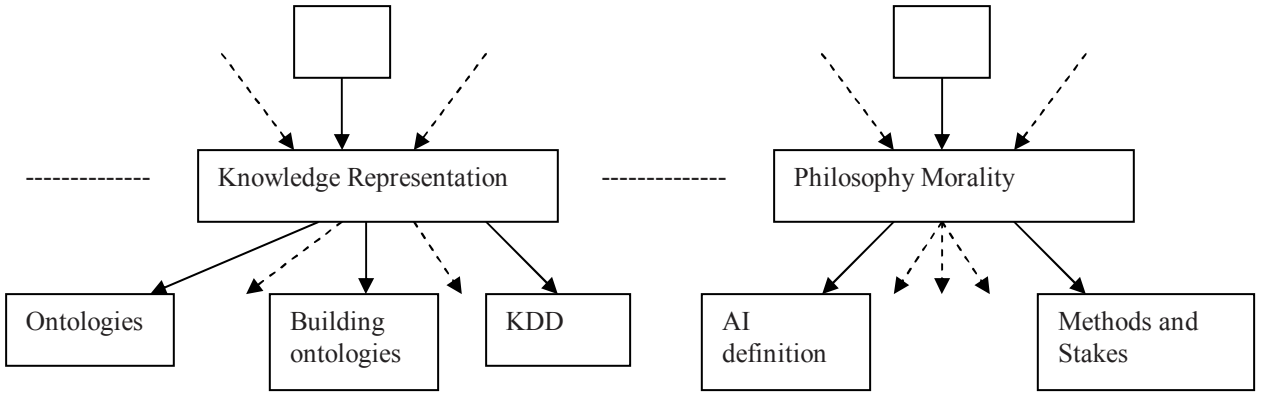Table 1 : extracted concepts by the algorithm presented in section 4.2

Figure 3: A part of the hierarchy automatically induced on LookSmart data

6.2.2. Similarity between Hierarchies

We have tested the following three methods on the two document collections (see section 4 for more details):
- The terms hierarchy of Croft obtained by the subsumption method
- Method 1: the concepts hierarchy is built on the term hierarchy of Croft
- Method 2: the subsumption definition is applied directly to the concepts, with the assignment of documents to concepts deduced from the estimation of *P(Concept | document)* via EM algorithm. For this method *P(concept$_1$|concept$_2$)* is estimated by counting.

For each concept hierarchy (words for 1., sets of keywords for 2. and 3.), each document collection is mapped onto the hierarchy and the two quality measures introduced in section 5 are computed.

	Croft Hierarchy			Method 1			Method 2		
	MI	MI$_D$	MI$_{P-C}$	MI	MI$_D$	MI$_{P-C}$	MI	MI$_D$	MI$_{P-C}$
LookSmart									
Mutual Information	0.3	0.5	0.1	0.6	0.7	0.5	0.7	0.8	0.6
NewScientist									
Mutual Information	0.2	0.3	0.1	0.2	0.4	0.0	0.67	0.7	0.64

Table 2: similarities between hierarchies built by the three tested methods and the originals ones.

If we compare the hierarchy of documents resulting from the term hierarchy of Croft with the original hierarchy of Looksmart, the similarity is low (0.3, column "Croft Hierarchy", table 2), although both hierarchies use single terms to index and organize the documents. Croft hierarchy uses most of the collection terms whereas Looksmart uses a much more restricted vocabulary. The former hierarchy is much larger and deeper than the original one. Note that in MI, the term which penalizes the similarity is MI$_{P-C}$ (corresponding to the detection of the "parent-child" relation). Remember $MI(X,Y) = \alpha * MI_D(X_D,Y_D) + (1 - \alpha) * MI_{P-C}(X_{P-C}, Y_{P-C})$, here $\alpha = 0.5$

The hierarchies obtained by our methods also have more nodes and are much deeper than the original hierarchies. This is due to the fact that certain topics discovered are not present in the original hierarchies which exploit a simple conceptual representation (single term for a concept). Nevertheless

the similarities are more significant, and they indicate a clear coherence between the induced and original hierarchies. This is not true for method 1 on the heterogeneous corpus NewScientist. This last phenomenon highlights the weakness of the term subsumption in the presence of heterogeneous data. Method 2 which directly computes subsumption between topics gives much better results.

Globally, the hierarchies obtained by organizing the documents on automatically extracted concepts are much closer to the original hierarchies than those built on the term hierarchy. These experiments highlight the behaviour of our algorithm.

6.2.3. Specialization /Generalization Property of Hierarchies

For this measure (section 5.2) of the generalization/ specialization quality of the hierarchies, the lower the index value is the better the method is.

	LookSmart	Croft Hierarchy	Method 1	Method 2
Specialization/ Generalization measure	41.53	20.62	15.2	3.8
	NewScientist	Croft Hierarchy	Method 1	Method 2
Specialization/ Generalization measure	50.12	45.2	32.11	10.87

Table 3 : Generalization / Specialization Quality

Results on table 3 show that the original hierarchies have a low generalization/ specialization quality according to the measure. The document organization produced by method 2 is clearly superior to the ones obtained with the other 2 methods which supports the idea of concept hierarchies as opposed to term hierarchies and highlights the validity of the proposed method.

7. Conclusions and Perspectives

We have described an automatic method based on the analysis of corpus statistics to infer "generalization/specialization" relations between concepts of a corpus. Other types of relations can be induced by the same kind of statistical analysis with additional information sources. The exploitation of the semantic relation "generalization/specialization" can lead to the generation of hierarchical structures of documents collection. We also introduced numerical measures for the open problem of the comparison and evaluation of such hierarchies. These measures are of two types: the first type gives an indication of similarity between hierarchies and allows to measure coherence among different hierarchies. This type of measure does not give an idea on the intrinsic quality of the hierarchies. The second type of measure quantifies the way in which a hierarchy respects the property of "generalization/specialization". Our method applied to the document collections extracted from the sites LookSmart and New-Scientist gives promising results which supports our idea that a hierarchical organization of collection can be generated automatically around discovered concept hierarchies. The experiments also show that our hierarchies of concepts are closer to the original hierarchies than those produced by a reference method which builds terms hierarchies automatically. Others experiments on various collections and larger corpora are necessary to confirm this fact.

Bibliographical References

J. Allan, 1996, Automatic hypertext link typing. *Proceeding of the ACM Hypertext*. Washington DC, USA, 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 Modeling 2001, LNAI 2109*.

M. Hearst, 1994, Multi-paragraph Segmentation of Expository Text. *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics*. Las Cruces.

M. Hearst, 1997, TextTitling : Segmenting Text into multi-paragraph Subtopic Passages. *Computational Linguistics*. pp. 33-64.

N. Hernandez, B. Grau, 2002. Analyse Thématique du Discours : segmentation, structuration, description et représentation.*CIDE'05*, Hammamet, Tunisie.

N. Hernandez, B. Grau, 2003 . What is this Text About ? *Proceedings of the 21st annual international conference on Documentation*. San Francisco, CA, USA*.*

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

J. Klavans, K. R. McKeown, M. Y. Kan, 1998, Ressources for Evaluation of Summarization Techniques. *In acts of First International Conference on Language Ressources & Evaluation (LREC)*. Grenade, Espagne, pp. 899-902.

D. Koller, M. Sahami, 1997, Hierarchically classifying documents usingvery few words. *In Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*.

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, W. Bruce Croft, 2000, Discovering and Comparing Topic Hierarchies. *Proceedings of RIAO conference*. pp 314-330.

D.Lawrie, B. Croft, A. Rosenberg, 2001, Finding Topic Words for Hierarchical Summarization. *Proceedings of the 24th annual international ACM SIGIR conference*. New Orleans, Louisiana, USA.

A. Maedche, S. Staab. Measuring Similarity between Ontologies. *European Conference on Knowledge Acquisition and Management, EKAW-2002, Madrid, spain, LNCS/LNAI2473, spinger 2002, pp 251-263*.

E. Mittendorf, P. Schäuble, 1994, Document and Passage Retrieval Based on Hidden Markov Models. *In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland, pp. 318-327.

E. Morin, 1999. Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. *Thèse en Informatique*, Université de Nantes.

E. Morin, C. Jacquemin, 1999. Expansion automatique de Thesaurus à partir de corpus. *Actes de la Troisième Conférence sur l'Ingénierie des Connaissances* (IC'99), Palaiseau, France, Juin 99, pp. 97-105

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

M. Sanderson, Bruce Croft, 1999, Deriving concept hierarchies from text. *In Proceedings ACM SIGIR Conference '99*. pp.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.

G. Ruge, 1997. Automatic Detection of Thesaurus relations for Information Retrieval. *Applications, Foundations of Computer Science: Potential - Theory - Cognition*, p.499-506.

I. Ryutaro, T. Hideaki, H. Shinichi. Rule Induction for Concept Hierarchy Alignment. In *Proceedings of the IJCAI-01 Workshop on Ontology Learning (OL-2001)*, pages 26–29, 2001.

A. Vinokourov, M. Girolami, 2002, 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.

W. Walker, 1991, Redundancy in collaborative dialogue. Actes of AAAI Symposium on Discourse Structure in Natural Language Understanding and Generation. Pacific Grove, USA.