Apprentissage de Relations « Généralisation / Spécialisation » entre Concepts – Application à la Structuration Hiérarchique Automatique de Corpus

Hermine Njike Fotzo, Patrick Gallinari

LIP6
8, rue du Capitaine Scott
75015 Paris
{Hermine.Njike-Fotzo, Patrick.Gallinari}@lip6.frr

RÉSUMÉ. *Nous étudions comment apprendre automatiquement à partir de corpus, des hiérarchies de concepts obéissant à une relation du type généralisation / spécialisation. Nous proposons une méthode qui permet à partir de concepts identifiés automatiquement sur un corpus de documents, d'apprendre des relations généralisation / spécialisation à partir de cooccurrence de ces concepts, puis de construire une hiérarchie ordonnée suivant cette même relation. A titre d'application, nous montrons comment utiliser cette hiérarchie de concepts pour construire une hiérarchie de documents. Nous introduisons des critères originaux qui permettent d'évaluer la qualité des hiérarchies ainsi construite et de les comparer entre elles ou avec des hiérarchies manuelles. Nous décrivons une série de tests réalisés sur des corpus de documents provenant de portails internet, ces corpus sont extraits des hiérarchies LookSmart et NewScientist.*

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 lead to the construction of the hierarchy. We resent 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 describe a series of tests performed on document collections coming from LookSmart and NewScientist hierarchies*

MOTS-CLÉS : *Structuration de corpus, relations sémantiques entre concepts, hiérarchies de concepts, segmentation de textes*

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

1. Introduction

La faible structuration de la majorité des collections de documents textuels disponibles limite la recherche d'information et la navigation. Certaines collections, comme celles des portails (Yahoo, Open Directory, LookSmart…), ont été manuellement structurées en des hiérarchies thématiques au prix d'efforts humains considérables. Dans ces hiérarchies, les documents sont classés dans des thèmes, qui sont eux-mêmes organisés en sous hiérarchies allant du plus général au plus spécifique (Källgren, 1988). L'utilisateur peut aisément naviguer dans ces hiérarchies. Au delà de la navigation, l'existence d'une structure de ce type sur une collection apporte une aide aux moteurs de recherche, facilite la maintenance et l'enrichissement des collections. Alors qu'il est simple de créer des hiérarchies de documents en fonction de leur similarité, en utilisant par exemple des méthodes de classification hiérarchiques, la construction de hiérarchies thématiques est beaucoup plus complexe.

Nous étudions comment apprendre automatiquement à partir de corpus de documents, des hiérarchies de concepts suivant une relation généralisation / spécialisation. Nous proposons une méthode qui à partir de concepts extraits d'un corpus permet d'apprendre une relation de généralisation / spécialisation entre ces concepts. Il est ensuite facile de créer une hiérarchie des concepts ainsi ordonnés.

Nous montrons à titre d'exemple comment utiliser cette hiérarchie de concepts pour construire une hiérarchie de documents. L'évaluation de la qualité de telles hiérarchies de documents est un problème ouvert et est largement subjective. Nous introduisons de nouveaux critères permettant le calcul d'indicateurs quantitatifs de la qualité des hiérarchies de documents ainsi construites. Nous testons ensuite les méthodes introduites sur deux corpus extraits de hiérarchies de sites internet.

Nous nous plaçons dans de cadre de méthodes complètement automatiques où toutes les connaissances sont apprises directement à partir du corpus. Il ne s'agit donc pas d'inférer des connaissances sémantiques fines au niveau de la phrase par exemple comme c'est souvent le cas en terminologie, ou de s'appuyer sur des ressources linguistiques spécifiques d'une thématique. Au contraire, il s'agit de construire des outils généraux capable d'extraire des relations sémantiques simples entre éléments d'un corpus, qui puissent être utilisées dans différentes tâches de recherche d'information.

L'article est organisé comme suit : la section 2 donne l'état de l'art, dans la section 3 nous introduisons les méthodes de base pour l'apprentissage de la relation de « généralisation/spécialisation ». La section 4 décrit en détail l'algorithme d'extraction des relations entre concepts du corpus. Dans la section 5 nous proposons des critères numériques pour mesurer la qualité des hiérarchies construites. Enfin la section 6 est consacrée aux expériences conduites sur les corpus extraits de deux sites : LookSmart et NewScientist.

2. Etat de l'art

Nous passons en revue un ensemble de travaux représentatif des recherches concernant la génération de hiérarchies de documents et la structuration de concepts. Nous mentionnons aussi des travaux sur la segmentation de textes en portions thématiquement homogènes sur lesquels nous nous appuyons pour extraire automatiquement les concepts des corpus.

2.2. Structuration de concepts et de collections de documents

Une méthode d'organisation des collections de documents qui a fait ses preuves en recherche d'information est le classement des documents au sein d'une hiérarchie de concepts qui sont eux-mêmes organisés en hiérarchie allant du plus général au plus spécifique. En recherche d'information, plusieurs approches ont été développées pour la génération des hiérarchies. Très souvent les hiérarchies sont créées manuellement, la seule partie automatique est alors le classement des documents au sein de la hiérarchie.

Les techniques de classification automatique ont souvent été utilisées pour créer des hiérarchies de documents. Ces techniques utilisent la similarité des documents mesurée le plus souvent sur des représentations vectorielles des fréquences de termes. Ce type de hiérarchie a été utilisé pour aider à la navigation et à la recherche d'information. Un exemple souvent cité est celui de Scatter/Gather (Cutting et al., 1992), l'algorithme regroupe de manière récursive les ensembles de documents pour créer la hiérarchie. Dans le même esprit mais à en utilisant un formalisme probabiliste, (Vinokourov et al., 2002), proposent un modèle qui permet d'inférer une structure hiérarchique pour l'organisation non supervisée d'une collection de documents. Les techniques de classification hiérarchique ont été largement utilisées pour organiser des corpus et aider ainsi à la recherche d'information. Toutes ces méthodes regroupent les documents en se basant uniquement sur leur similarité. A chaque niveau de la hiérarchie on a des regroupement de plus en plus gros fusionnant les précédents groupes suivant leurs similarités. Dans ce type de hiérarchies, il n'y a pas de relation sémantique entre les nœuds de différents niveaux, par la nature même de leur construction. Elles ne peuvent pas être utilisées pour inférer des relations sémantiques nommées entre les concepts représentés par chaque regroupement. En particulier ces organisations hiérarchiques sont de peu d'utilité pour naviguer des collections. Il est difficile avec ces méthodes d'expliquer le contenu de chaque niveau de la hiérarchie et de les interpréter.

Récemment, de nouveaux types de hiérarchies construites automatiquement ont été proposés. Ce sont des hiérarchies des termes qui apparaissent dans un ensemble de documents. Elles sont bâties à partir de relations de généralisation/spécialisation entre termes découvertes automatiquement sur le corpus exploité. On peut en particulier citer les hiérarchies construites en utilisant la relation de subsomption entre termes de (Lawrie et al., 2000), (Sanderson et al., 1999) qui sera présentée dans la section 3. Une fois ces hiérarchies de terme construites, il est possible de « projeter » les documents du corpus sur la hiérarchie de termes et d'offrir ainsi un résumé assez complet de l'ensemble des documents qui se prête bien par son

organisation à la navigation par exemple. Dans la même veine, (Krishna et al., 2001) proposent un cadre général de modélisation pour des relations asymétrique entre données.

Par rapport à ces travaux, nous proposons une contribution originale. Il s'agit de l'extension de ce type d'approche à la construction de hiérarchies de véritables concepts (thèmes) représentés par un ensemble de mots clés et non plus par un seul terme. Ceci est réalisé par la détection automatique de relations de « généralisation/spécialisation » entre ces concepts qui sont découverts automatiquement à partir du corpus. Ceci permet en particulier de naviguer dans une collection en se basant sur les sujets traités et non seulement sur les mots qui y sont présents. A partir de cette hiérarchie de concepts, on peut générer des liens de « généralisation/spécialisation » entre documents qui peut être un outil supplémentaire de navigation entre les documents.

Dans la communauté du langage naturel, il y a de nombreux travaux portant sur la détection semi-automatique de relations entre termes. Les travaux sur la terminologie sont au centre de ces problèmes. On peut citer en particulier l'étude du comportement des termes en discours, la construction des dictionnaires et les référentiels terminologiques dans un domaine qui renvoient aux problématiques de l'automatisation du repérage et de la structuration des termes, la détection des relations sémantiques partagés par les termes (Morin, 1999), (Morin et al., 1999) (Ruge et al., 1997). Les relations trouvées sont ensuite validées par un humain. Ces relations opèrent au niveau de la sémantique de la phrase et à un niveau beaucoup plus fin que celui auquel nous nous plaçons. Par opposition, nous recherchons des relations plus générales non pas au niveau des phrases mais entre parties de documents et nous cherchons à automatiser totalement le processus.

(Hernandez et al., 2002), (Hernandez et al., 2003] s'intéressent à la définition de principes facilitant l'accès au contenu d'un seul document. La philosophie utilisée par les auteurs est assez proche de la notre. Ils se basent sur la segmentation thématique pour mettre en évidence la structure du texte et décrire les segments par les descriptions de leurs thèmes et l'identification de leurs rôles. La description des thèmes se fait à deux niveaux : global et local. Les différents segments thématiques sont représentés par des groupes nominaux. L'importance thématique d'un groupe nominal est calculée en fonction de sa fréquence (tf-idf) au niveau local par rapport à chaque segment thématique et au niveau global par rapport au document entier. Cette structuration permet entre autres, la navigation intra-documents, le reformatage dynamique des documents électroniques qui ont pour la plupart une mise en page papier pour une visualisation ciblée. La différence entre ces travaux et les notres est que nous travaillons au niveau de la structuration d'une collection et non pas d'un seul document. Nous nous intéressons aux relations entre thèmes d'un corpus et non entre différentes parties d'un document.

2.2. Segmentation de textes

Avec l'augmentation de la taille et de la complexité des documents traités, la recherche et l'extraction de passages pertinents ont connu un essor particulier depuis une dizaine d'années. Découper un texte, en passages cohérents apporte une

information substantielle pour de nombreuses applications en recherche d'information (Kalvans et al,. 1998). La tâche de segmentation consiste à identifier des régions de texte possédant des propriétés pertinentes pour une tâche donnée. Ici nous nous intéressons à la segmentation de document en passages cohérents et homogènes (Hearst et al., 1997). En particulier, nous avons choisi d'utiliser la technique de segmentation thématique de (Salton et al., 1996). Elle utilise l'hypothèse que si les représentations vectorielles de deux extraits ont une faible similarité alors ces extraits ont de faibles liens thématiques. Ainsi deux extraits peu similaires (au sens d'une mesure donnée) donneront lieu à une segmentation du document qui les contient. L'algorithme de (Salton, 1996) procède à la décomposition des textes en segments et en thèmes où un segment est un bloc de texte contigu traitant d'un seul sujet et un thème est un ensemble de tels segments. Dans cette approche, le processus de segmentation commence au niveau des paragraphes. Ce choix d'unité minimale peut se justifier par le fait que les auteurs d'un texte exposent en général un point de vue par paragraphe. Plus précisément la méthode de Salton et al. pour la décomposition d'un document en thèmes s'articule autour des points suivants :

- Calculer les similarités entre différents paragraphes du document et retenir celles qui sont supérieures à un certain seuil1. Construire le graphe de similarité et en extraire les triangles. Un triangle est un ensemble de trois paragraphes fortement liés les uns aux autres, et donc susceptible de représenter une thématique cohérente.
- Pour chaque triangle construire un vecteur centroïde (représentation vectorielle) qui est la moyenne des trois vecteurs représentant les paragraphes du triangle.
- Fusionner les triangles dont la similarité des vecteurs centroïdes est supérieure à un seuil2. Réitérer la fusion jusqu'à satisfaction d'un critère de convergence

3. Relation de « Généralisation/Spécialisation »

Entre deux entités D1 et D2, il existe une relation de type « généralisation/spécialisation » (par exemple D2 est une spécialisation de D1 et D1 une généralisation de D2) si D2 évoque une spécificité de D1, ou traite de concepts spécifiques aux concepts traités dans D1. Par exemple D1 = sport et D2 = football, ou alors D1 est un document qui traite du thème de la guerre en général et D2 traite de la première guerre mondiale en particulier. Ce type de relation permet de construire une organisation hiérarchique de concepts présents dans un corpus et d'en dériver une organisation hiérarchique.

D'autres types de relations peuvent conduire à une organisation hiérarchique de concepts et de documents d'une collection (par exemple la relation de pré-requis), mais la relation de « généralisation/spécialisation » ou « est-un » est la plus répandue pour l'organisation des collections, sans doute aussi parce qu'elle est la plus intuitive pour un utilisateur qui veut exploiter une collection. Cette relation est utilisée pour l'organisation des données sur des portails tels Yahoo, Open Directory,

LookSmart… C'est aussi une des relations prépondérantes dans les ontologies avec l'hyponymie, la co-hyponymie, la synonymie…

3.1. Détection de la relation de « généralisation/spécialisation » entre deux concepts

Dans la plupart des organisations hiérarchiques de documents actuelles les concepts sont réduits aux mots. En général, la hiérarchie des concepts est construite manuellement et c'est l'affectation des documents aux noeuds de la hiérarchie qui est automatique. Néanmoins en 1999, (Croft et al., 1999) proposent une définition de la généralité d'un mot par rapport à un autre basée sur les statistiques des mots dans la collection. La notion de généralité/spécificité est basée sur la subsomption entre termes. En effet, étant donné une collection de documents, certains termes apparaissent fréquemment dans l'ensemble des documents et d'autres n'apparaissent que dans peu de documents. Certains des termes fréquents donnent un ensemble d'information important sur les thèmes abordés dans les documents. Certains termes définissent de manière générale un sujet, alors que d'autres qui co-occurrent avec ces termes généraux expliquent des aspects du sujet. La subsomption tente de mettre en avant les caractéristiques des différents concepts et leurs relations.

L'idée clé de Croft et Sanderson, est d'avoir utilisé une mesure simple pour caractériser la subsomption. La subsomption caractérise une relation de généralité/spécificité entre deux termes et est basée sur un principe de co-occurrence non symétrique :

Le terme x subsume (ou est plus général que) le terme y $\Leftrightarrow P(x|y) > t$ *et* $P(y|x) < P(x|y)$*, où t est un seuil fixé.*
n(x,y) = est le nombre de documents contenant les termes x et y.
n(y) = est le nombre de documents contenant le terme y.
$P(x|y) = n(x,y) / n(y)$.

En d'autres termes, x subsume y si les documents où y est présent sont un sous ensemble ou proche d'un sous ensemble des documents contenant x. la seconde règle assure que si les deux co-occurrent plus de t% fois, le terme le plus fréquent va être considéré comme le plus général.

Cette mesure qui est proposée à la base pour la subsomption des termes peut s'étendre à la subsomption des thèmes (ou sujets), où chaque thème est représenté par un ensemble de mots clés ainsi que nous le verrons dans la section 4.

3.2. Remarques sur la mesure de subsomption et sur d'autres exploitations de ces statistiques

La première remarque importante concernant cette mesure de subsomption, est qu'elle n'est appropriée que dans les domaines ou il y a souvent répétition de termes dans le langage. Dans le cas contraire les estimations de co-occurrences ne seraient pas pertinentes. Cette définition sera donc valable pour les corpus scientifiques et moins pour les corpus littéraires ou les articles de journaux. Néanmoins, en incorporant des ressources linguistiques comme WordNet pour prendre en compte

les synonymies, on peut réduire la sensibilité de la technique à la variabilité dans les corpus. D'autres part, on peut pressentir que cette définition générera des relations pertinentes sur des corpus homogènes où tous les documents traitent du même sujet. Avec cette définition de la subsomption, un concept peut avoir plusieurs parents. Les différents chemins d'accès à ce concept correspondent aux différents sens du concept et reflètent donc sa polysémie.

Dans ce papier nous nous focalisons sur la relation de « généralisation/spécialisation » que nous étendons aux thèmes. Néanmoins, d'autres types de relations peuvent être extraits avec l'exploitation de statistiques similaires, mais nécessitent des sources d'information supplémentaires pour nommer ces relations. Par exemple, les collocations au sein d'une phrase vont souvent modéliser des relations symétriques (synonyme, co-hyponyme, …). De même les collocations du type prochain-voisin modélisent généralement les relations anti-symétriques comme l'hyponymie, les relations propriété-possesseur ou classe-instance… Les réseaux de collocations peuvent donner l'intuition d'une relation entre deux entités, mais ne permettent pas de la nommer sans autre source d'information ou analyses supplémentaires.

4. Extraction de Relations et Génération de Hiérarchies

Pour la génération de relation de « généralisation/spécialisation » entre concepts et la génération des hiérarchies de ces derniers, on commence par extraire les concepts (ou thématiques) présents dans les documents du corpus. Contrairement aux hiérarchies actuelles où les concepts sont réduits aux termes on cherche ici à construire des hiérarchies de thèmes représentés par un ensemble de mots clés. En effet, les mots ne donnent qu'une indication très rudimentaire du contenu d'un document et les hiérarchies correspondantes sont assez pauvres.

4.1. Prétraitement et représentation des documents

Notons $V = \{w_j\}_{j\in\{1,\ldots,M\}}$ l'ensemble des mots du vocabulaire ne comportant que les mots radicalisés et excluant les mots courants de la langue (qui n'apportent pas d'information) et les mots rares (apparaissant dans moins de 5 documents ou x% des documents de la collection). $D = \{D_i\}_{i\in\{1,\ldots,N\}}$ est l'ensemble des documents de la collection, $P=\{P_k\}_{k\in\{1,\ldots,L\}}$, l'ensemble des paragraphes contenus dans l'ensemble des documents. Les paragraphes vont nous servir comme unité de base pour la détection des thèmes présents dans le corpus.

Un document D_i sera représenté par un vecteur de caractéristiques comme suit :

$Di = (tf_i(w_1)*idf(w_1), \ldots, tf_i(w_M)*idf(w_M))$,

où $tf_i(w_j)$ est la fréquence du terme j dans le document D_i, $idf(w_j) = \log(N/df(w_j))$, avec N le nombre de documents de la collection et $df(w_j)$ le nombre de documents de la collection contenant w_j. C'est la représentation classique des documents en sac de mots.

De même un paragraphe sera représenté comme suit :

$P_k = (tf_k(w_1)*ipf(w_1), \ldots, tf_k(w_M)*ipf(w_M))$,

où $tf_k(w_j)$ est la fréquence du terme j dans le paragraphe P_k. $ipf(w_j) = \log(L/dp(w_j))$, avec L le nombre de paragraphes de la collection et $dp(w_j)$ le nombre de paragraphes de la collection contenant wj.

La mesure de similarité utilisée entre deux entités (documents ou paragraphes) est le cosinus entre leurs vecteurs caractéristiques.

4.2. Extraction des concepts du corpus

Le but est d'extraire l'ensemble des concepts d'un corpus et les mots qui les représentent le mieux. Pour cela nous étendons la méthode de segmentation de (Salton et al., 1996) : dans un premier temps, nous décomposons, chaque document en un ensemble de thèmes sémantiques avec la méthode de Salton. Ensuite, nous procédons à un regroupement des thèmes pour retenir l'ensemble minimal de thèmes couvrant le corpus :

- On construit un graphe de similarité entre les thèmes de tous les documents (il existe un arc entre deux thèmes si leur similarité est supérieure à un certain seuil)
- Ensuite on recherche les composantes connexes de tous les thèmes. Dans chacune des composantes, on ne retient que les nœuds qui ont une liaison dans le graphe avec au moins 75% des autres nœuds de la composante
- On fusionne les composantes qui ont une relation d'inclusion (stricte ou à β%)
- Les composantes restantes forment l'ensemble des thématiques du corpus

Chaque thème est représenté par un ensemble de mots qui le représente le mieux. Dans nos tests ce sont les mots les plus fréquents. A partir de maintenant nous identifierons la notion de concepts à ces thématiques représentées par des mots clés.

Algorithme EM :

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

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

$P_d = P(d \in C) = P(C \mid d)$

$P_t = P(t \in C) = P(C \mid t)$

Initialisations : initialiser P_t^C avec la connaissance des mots clés des concepts

P_t^C = # de termes t dans le concept C / # de termes dans le concept C

Étape E: $P_t = P(t \mid C) * P(C) / P(t) = P_t^C * P(C) / P(t)$

$P_d = [1/P(C)]*\prod_{t \in d} P(C \mid t) = [P(C)/P(d)] * \prod_{t \in d} [P(C \mid t)P(t) / P(C)]$

Étape M: ré-estimation de P_t^C avec les résultats de l'étape E

P_t^C = # de documents $\in$ C possédant le terme t / # documents $\in$ C

$d \in T \Leftrightarrow Pd >$ seuil

Log Vraisemblance :

$V = P(D \mid \Theta) = \prod_d P(d \mid \Theta) = \prod_d \prod_{t \in d} P(t \mid \Theta)$

$= \prod_d \prod_{t \in d} \sum_c P(t, C \mid \Theta) = \prod_d \prod_{t \in d} \sum_c P(C \mid t, \Theta) P(t \mid \Theta)$

$Log(V) = \sum_d \sum_{t \in d} \log(\sum_c P(C \mid t, \Theta) P(t \mid \Theta))$

D désigne le corpus et Θ les paramètres du modèle.

Figure1 : algorithme EM pour l'estimation de P(Concept|document)

4.3. Induction des relations de « généralisation/spécialisation » entre concepts

Nous allons introduire deux méthodes pour la construction de hiérarchies de concepts. La première s'appuie sur la hiérarchie de termes de l'algorithme de Croft et al., l'expérience nous a montré qu'elle avait des défauts. Nous proposons ensuite une deuxième méthode qui infère automatiquement des relations entre thèmes sans passer par une hiérarchie de termes.

4.3.1. Méthode1 : exploitation de la hiérarchie de termes de Croft et al.

La première méthode que nous proposons détecte les relations entre concepts par l'exploitation de la hiérarchie de termes (Croft et al., 1999) construite sur le corpus. La hiérarchie de concepts est construite comme suit : pour chaque couple de concepts (C_1, C_2), on calcule à partir de la hiérarchie de termes le pourcentage *x* de mots de C_2 généralisés par les mots de C_1 et *y* le pourcentage de mots de C_1 généralisés par les mots de C_2 si $x > S_1 > S_2 > y$ (S_1 et S_2 sont des seuils), on en déduit une relation de « généralisation/spécialisation » entre ces concepts (C_1 généralise C_2).

Cette méthode hérite des faiblesses de celle proposée de Croft et al. En particulier, elle ne fonctionne que sur des corpus très homogènes avec forte répétition de termes. Afin de s'affranchir de la hiérarchie de termes issue de la méthode de Croft et al, nous proposons une deuxième méthode.

4.3.2. Méthode2 : application directe de la définition de subsomption aux concepts

La seconde approche consiste à estimer directement les probabilités conditionnelles d'un concept sachant un autre $P(C_i|C_j)$ sans passer par une décomposition en mots. L'estimation de ces probabilités pour toute paire de concepts permet d'appliquer la définition de subsomption directement aux concepts. $P(C_i|C_j)$ peut être estimé par comptage :

$P(C_i|C_j)$ =(nombre de documents traitant de C_i et C_j)/(nombre de documents traitant de C_j)

La difficulté réside dans le calcul des $P(C|d)$ qui détermine l'affectation d'un concept à un document qui est nécessaire pour calculer les $P(C|C_j)$, ce que nous commentons ci-dessous.

Les thèmes sont déterminés par segmentation de paragraphes (cf section 4.2). Ils sont identifiés par les mots les plus fréquents. Les résultats de la segmentation peuvent être utilisés directement pour affecter les thèmes aux documents. En effet, a l'issue de cette segmentation on connaît les documents d'où proviennent les paragraphes composant le thème. Ce moyen d'affection des thèmes aux documents donne des estimations rudimentaires de $P(C|d)$ et d'autre part, de nombreux documents qui parlent du concept mais qui ne possèdent pas un paragraphe entier associé à celui-ci vont être ignorés. Nous proposons plutôt de procéder à l' estimation $P(C|d)$ via un algorithme EM simple qui est donnée ci-dessus.

Une fois les relations de « généralisation/spécialisation » détectées sur les couples de concepts, on applique la transitivité pour terminer la construction de la hiérarchie des concepts.

Après la construction de la hiérarchie de concepts, on peut indexer les documents par les thèmes qu'ils contiennent et les affecter aux différents nœuds. Un document peut appartenir à plusieurs noeuds s'il traite de plusieurs thèmes.

5. Mesures d'évaluation

Evaluer la pertinence d'une hiérarchie est un challenge et reste un problème largement ouvert. Les évaluations faites sur des groupes d'utilisateurs donnent généralement des résultats ambigus et partiels. Les mesures automatiques elles, ne donnent que des intuitions sur la valeur intrinsèque des hiérarchies. Néanmoins, à ce stade pour s'affranchir du lourd processus d'une évaluation humaine, on a recourt à des critères automatiques pour juger la pertinence des hiérarchies apprises et par là même la pertinence des relations détectées entre concepts. Pour évaluer notre approche, nous proposons deux mesures complémentaires. La première est un indicateur de similarité entre hiérarchies. Cela nous permettra par exemple de comparer la cohérence de nos hiérarchies automatiques par rapport à des hiérarchies manuelles sans pour autant fournir d'indicateur de qualité de l'une ou de l'autre de ces hiérarchies. La seconde reflète à quel point une hiérarchie respecte le caractère de généralisation/spécialisation entre les éléments de ses nœuds.

5.1. Mesure de similarité entre hiérarchies de documents

Les documents d'une hiérarchie partagent une relation « frère » s'ils appartiennent au même nœud et une relation « parent-descendant » s'ils appartiennent à la même branche. La mesure de similarité que nous proposons est basée sur l'information mutuelle entre hiérarchies et s'inspire de la mesure de similarité pour comparer deux algorithmes de clustering proposée dans (Draier et al., 2001). Soient X et Y les étiquettes (classes) de tous les éléments de l'ensemble des données correspondant respectivement à deux algorithmes de clustering, X_i l'étiquette du i^{ème} cluster pour X, $P_X(C = k)$ la probabilité qu'un objet appartienne au cluster k dans X, et $P_{XY}(C_X=k_x, C_Y=k_y)$ la probabilité jointe qu'un objet appartienne au cluster k_x dans X et au cluster k_y dans Y. Pour mesurer la similarité des deux méthodes de clustering, les auteurs proposent d'utiliser l'information mutuelle entre les deux distributions de probabilités:

$MI(X,Y) = \Sigma_{i \in CX} \Sigma_{j \in CY} P_{XY}(C_X = i, C_Y = j) * log [(P_{XY}(C_X = i, C_Y = j)) / (P_X(C_X = i) * P_Y(C_Y = j))]$. Si MI est normalisée entre 0 et 1, plus $MI(X, Y)$ est proche de 1 plus les deux ensemble de clusters sont similaires et les méthodes aussi.

Dans le cas d'organisations hiérarchiques, pour mesurer la similarité entre deux hiérarchies, nous devons mesurer la façon dont les documents sont regroupés ensemble dans les mêmes noeuds et également la similarité des relations "parent-descendant" entre documents dans les deux hiérarchies. Pour simplifier, nous allons dans un premier temps considérer que chaque document n'appartient qu'à un cluster unique. L'extension au cas où il apparaîtrait dans différents noeuds est aisée et n'est pas exposée ici.

Pour une hiérarchie X notons X_i un nœud de la hiérarchie. Une hiérarchie de documents est décrite par deux relations qui sont la relation « frère » que partagent les documents au sein d'un nœud et la relation de généralisation entre couples de documents partageant une relation de descendance (relation « parent-descendant »). Une hiérarchie peut donc être vue comme issue de deux regroupements simultanés portant respectivement sur les documents et sur les couples « parent-descendant ». Elle est définie par les groupes de documents liés par ces deux relations.

L'information mutuelle $MI(X,Y)$ entre deux hiérarchies sera la combinaison de deux composantes : $MI_D(X_D,Y_D)$ l'information mutuelle entre les groupes de documents correspondant aux nœuds des deux hiérarchies (c'est la même mesure que pour un clustering classique) et $MI_{P\text{-}C}(X_{P\text{-}C}, Y_{P\text{-}C})$ l'information mutuelle mesurée sur les groupes de couples « parent-descendant » des hiérarchies. L'information mutuelle entre les hiérarchies X et Y sera alors calculée par :

$MI(X,Y) = \alpha * MI_D(X_D,Y_D) + (1 - \alpha) * MI_{P\text{-}C}(X_{P\text{-}C}, Y_{P\text{-}C})$, où α est un paramètre qui permet de donner plus ou moins d'importance au regroupement de documents dans les mêmes nœud ou aux relations hiérarchiques « parent-descendant » entre documents.

Cette mesure permet de comparer des hiérarchies de structures différentes. Elle permet d'analyser les similarités entre hiérarchies. Notamment, les résultats des

différents termes de la mesure nous permet de savoir ce qui contribue le plus à la similarité : regroupement des documents ou relations de descendance entre documents.

La mesure présente néanmoins quelques limites. Elle ne prend pas en compte la profondeur dans la relation de dépendance. De plus les relations frères ne sont considérées que par paire et non de façon plus globale. On peut avoir des hiérarchies semblables selon la mesure alors qu'elles ont des intérêts bien différents. Il suffit par exemple de prendre un nœud d'une hiérarchie et de le scinder en autant de nœuds que de paires de documents qu'il contient. La similarité avec la première hiérarchie sera de 1 pourtant la première est plus synthétique et sûrement plus intuitive.

5.2. Quantifier la capacité de « généralisation/spécialisation » d'une hiérarchie

La deuxième mesure que nous proposons quantifie la capacité d'une hiérarchie à respecter la relation de généralisation / spécialisation entre les objets qui sont dans les nœuds. Elle est basée sur l'entropie conditionnelle.

L'entropie conditionnelle mesure l'incertitude qu'on a sur une variable sachant une autre : $H(Y|X)= -\Sigma_x \Sigma_y P(x, y) * log(P(y|x))$. Dans le cadre de la subsomption, si un terme x généralise un terme y alors l'incertitude sur x sachant y est faible. On note IGT l'Indice de Généralisation d'un Terme et IGH l'Indice de Généralisation d'une Hiérarchie :

$IGT (t, \{f_i\}) = \Sigma_{fi} - P(t, f_i) * log(P(t|f_i))$, plus cet index sera faible plus *t* sera un bon généralisant de ses fils $\{f_i\}$

$IGH = \Sigma_{noeud} IGT(noeud, \{fils\}_{noeud})$.

Cette mesure conçue pour des hiérarchies de termes s'étend facilement aux hiérarchies de thèmes. Nous ne détaillerons pas cette extension ici.

6. Expériences et Résultats

6.1. Données

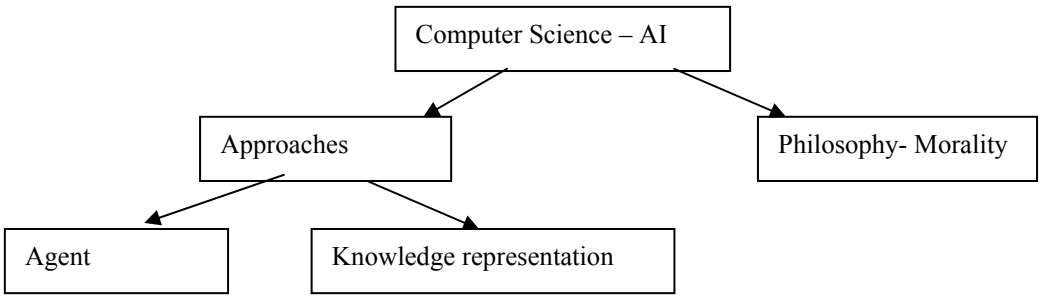

Figure2: Exemple de hiérarchie - Sous hiérarchie du site LookSmart utilisée dans les expériences

Les données utilisées pour les expériences sont des sous-hiérarchies des sites www.looksmart.com et www.newscientist.com. La sous-hiérarchie de Looksmart est composée de 100 documents et d'environ 7000 termes, celle de NewScientist est composée de 700 documents et environ 20000 termes. Le site NewScientist est un magazine hebdomadaire sur la science et la technologie, il contient les dernières informations de ces domaines. La hiérarchie extraite de ce site est hétérogène. Contrairement aux données de Looksmart qui ne traite que de l'intelligence artificielle, ici les documents traitent de l'intelligence artificielle, du clonage, du bioterrorisme, des dinosaures et de l'Iraq. Pour chacune des hiérarchies extraites, on a pour chaque thème des sous-catégories concernant des aspects spécifiques du thème.

On compare les hiérarchies induites par nos méthodes aux hiérarchies originales sur les mêmes données en utilisant les mesures proposées dans la section 5.

6.2. Résultats des expériences

6.2.1. Exemples de concepts et de relations extraits

	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 : concepts extraits par l'algorithme présenté en section 4.2

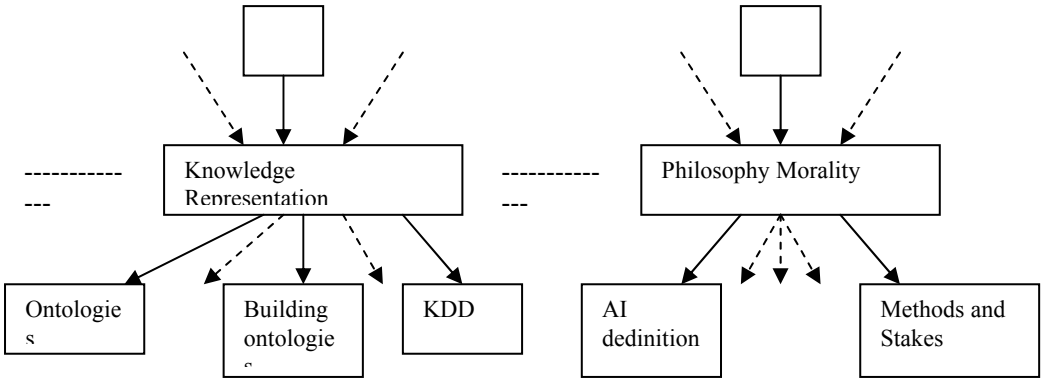

Figure3: Une partie de la hiérarchie de concepts induites automatiquement sur les données Looksmart.

Dans la table ci-dessus figurent quelques exemples de concepts extraits sur le corpus LookSmart et leurs relations. Chaque concept est identifié par un ensemble de mot clés (qui sont ici les mots les plus fréquent du thème). L'algorithme a découvert des relations de « généralisation/spécialisation » entre les concepts (2,3), (2,4), (2,5).

Par comparaison avec la hiérarchie Looksmart qui a cinq catégories au départ, la hiérarchie induite par notre algorithme sur les mêmes données est plus large et plus profonde. Les catégories sont plus spécifiques et l'algorithme découvre plus de thématiques. Par exemple, plusieurs catégories émergent de la catégorie « Knowledge Representation » de départ : ontologies, building ontologies, KDD (où les papiers parlent des représentation de données pour KDD) … de même la catégorie « Philosophy-Morality » est subdivisée en plusieurs catégories comme AI definition, Methods and stakes…

6.2.2. Similarité entre Hiérarchies

Nous avons testés les trois méthodes sur les deux corpus de documents. Pour chaque hiérarchie de concepts construite, les documents y sont projetés. Les mesures de qualité portent sur ces hiérarchies de documents.

Les méthodes testées sont les suivantes (voir la section 4 pour plus de détails):

- La hiérarchie de Croft est la hiérarchie de termes obtenue par la méthode de subsomption
- Méthode 1 c'est la hiérarchie de concepts construite à partir de la hiérarchie de termes de Croft
- Méthode 2 c'est l'application directe de la définition de subsomption aux concepts, avec l'affectation des documents aux concepts issue de l'estimation de *P(Concept | document)* par l'algorithme EM. Pour cette méthode $P(concept_1|concept_2)$ est estimé par comptage.

	Hiérarchie Croft			**Méthode 1**			**Méthode 2**		
	MI	MI_D	MI_{P-C}	MI	MI_D	MI_{P-C}	MI	MI_D	MI_{P-C}
LookSmart									
Information Mutuelle	0.3	0.5	0.1	0.6	0.7	0.5	**0.7**	0.8	0.6
NewScientist									
Information Mutuelle	0.2	0.3	0.1	0.2	0.4	0.0	**0.67**	0.7	0.64

Table 2 : similarités entre hiérarchies construites par les trois méthodes testées et la hiérarchie originale.

Si nous comparons la hiérarchie de document issue de la hiérarchie de termes de Croft avec celle d'origine sur la base Looksmart, la similarité est faible (**0.3,** colonne « Hiérarchie Croft » table 2), bien que les deux hiérarchies utilisent les termes pour indexer et organiser les documents. En fait, la hiérarchie Croft utilise la plupart des termes de la collection alors que Looksmart utilise un vocabulaire beaucoup plus restreint. La hiérarchie basée sur les termes est donc beaucoup plus large et plus profonde que celle d'origine mais les regroupements de documents restent cohérents dans les deux hiérarchies. En effet, le terme qui pénalise la

similarité est MI_{P-C} (correspondant à la détection de la relation « parent-descendant ».

On rappelle que $MI(X,Y) = \alpha * MI_D(X_D,Y_D) + (1 - \alpha) * MI_{P-C}(X_{P-C}, Y_{P-C})$, ici $\alpha = 0.5$)

Les hiérarchies obtenues par nos méthodes en organisant les documents par rapport à la hiérarchie de concepts découverts sur le corpus, possèdent également plus de nœuds et sont plus profondes que les hiérarchies originales. Ceci est dû au fait que certains thèmes découverts ne sont pas présent dans les hiérarchies originales qui exploitent une représentation conceptuelle simple (terme unique pour un concept). Néanmoins les similarités sont plus satisfaisantes, et elles dénotent la cohérence entre les hiérarchies induites et originales. Ceci est moins vrai pour la méthode 1 et le corpus hétérogène NewScientist. Ce dernier phénomène met en évidence la faiblesse de la subsomption des termes en présence de données hétérogènes. La méthode 2 qui utilise directement la subsomption des thèmes sans passer par les termes donne de bien meilleurs résultats.

Globalement, les hiérarchies obtenues en organisant les documents par rapport aux concepts extraits automatiquement sont plus proches des hiérarchies originales que celles construites sur la hiérarchie des termes. Ces expériences éclairent le comportement de notre algorithme. A l'examen, les hiérarchies induites sont cohérentes avec les hiérarchies extraites des portails.

6.2.3. Propriété de « généralisation/spécialisation » des hiérarchies

Pour cette mesure (section 5.2) de capacité de généralisation, plus la valeur de l'indice est faible, meilleure est la méthode.

	LookSmart	**Hiérarchie Croft**	**Méthode 1**	**Méthode 2**
Mesure spécialisation / généralisation	41.53	20.62	15.2	**3.8**
	NewScientist	**Hiérarchie Croft**	**Méthode 1**	**Méthode 2**
Mesure spécialisation / généralisation	50.12	45.2	32.11	**10.87**

Table 3 : Capacité de généralisation / spécialisation

Les résultats de la table ci-dessus montrent que les hiérarchies originales ont une faible capacité de généralisation. L'organisation produite par la méthode 2 semble être la meilleure au regard de cette mesure de capacité de généralisation. Néanmoins ces résultats intéressants doivent encore être pris avec précaution car ce qui sera vraiment pertinent sera la manière dont les utilisateurs appréhenderont les différentes hiérarchies.

7. Conclusions et Perspectives

Nous avons décrit une méthode automatique basée sur l'analyse des statistiques d'un corpus pour inférer des relations de « généralisation/spécialisation » entre concepts de ce corpus. D'autres types de relations peuvent être inférés par ce même type d'analyses statistiques moyennant des sources d'information supplémentaires pour les nommer. L'exploitation de la relation sémantique « spécialisation/généralisation » peut conduire à la génération d'une structuration hiérarchique d'une collection de documents. Nous avons également introduit des mesures numériques pour le problème ouvert de la comparaison et l'évaluation des tels hiérarchies. Ces mesures sont de deux types : un premier type donne l'indication de proximité entre hiérarchies et permet de mesurer la cohérence entre différentes hiérarchies. Ce type de mesure ne donne pas d'idée sur la qualité intrinsèque des hiérarchies. Le deuxième type de mesure quantifie la façon dont une hiérarchie respecte la propriété de « généralisation/ spécialisation ». Notre méthode appliquée aux données des sites LookSmart et New-Scientist donne des résultats intéressants et nous conforte dans l'idée que les organisations hiérarchiques de collection peuvent être générées automatiquement. Les expériences montrent aussi que nos hiérarchies de concepts sont plus proches des hiérarchies originales que celles produites par une méthode de référence qui construit automatiquement les hiérarchies de termes. D'autres expériences sur différentes collections et sur des corpus plus volumineux sont nécessaires pour confirmer ce fait.

Nos travaux ouvrent sur plusieurs perspectives dont les plus immédiates sont :

- L'algorithme EM conduit à l'estimation des probabilités P(C|terme) qui peuvent être utilisés comme alternative à la représentation des thèmes. Il serait intéressant de comparer ces deux méthodes de représentation de concepts.
- L'estimation de $P(C_i|C_j)$ avec les données P(C|document) et P(C|terme). Quelle influence sur l'organisation hiérarchique?
- L'exploitation de la méthode pour les liens de type « généralisation/spécialisation » entre documents

Bibliographie

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, 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.

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.

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.

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.