

Apprentissage d'une hiérarchie de concepts pour la conception de modèles de domaine d'hypermédias

Hermine Njike Fotzo, Thierry Artières, Patrick Gallinari, Julien Blanchard, Guillaume Letellier
LIP6, Université Paris 6
8 rue du capitaine Scott, 75015, Paris, France
{Prénom.Nom}@lip6.fr

Résumé. Nous décrivons comment apprendre automatiquement une hiérarchie de concepts à partir d'une collection de documents. Les concepts, identifiés par des ensembles de mots-clés, sont organisés en une hiérarchie de type spécialisation/généralisation. Cette hiérarchie peut être utilisée pour construire un modèle de domaine pour des collections de documents hypermédias. Nous proposons des idées sur la façon de construire des modèles d'utilisateurs à partir de tels modèles de domaine. Les modèles d'utilisateurs et de domaine peuvent être visualisés à l'aide d'outils efficaces comme les *Treemaps*.

1 Introduction

Un hypermédia adaptatif est personnalisé, dynamiquement, en fonction de l'utilisateur. La personnalisation peut consister en une adaptation du contenu de l'hypermédia ou à des aides à la navigation en ajoutant/enlevant des liens (Brusilovsky 1996). La personnalisation repose sur un modèle de l'utilisateur. Il n'existe pas aujourd'hui de consensus sur la définition d'un modèle d'utilisateur, souvent défini de façon ad-hoc, mis à part pour les hypermédias éducatifs et les systèmes tutoriaux (Da Silva 1998, De Bra 2003, Henze 1999). Ces systèmes utilisent un modèle de domaine, conçu manuellement, à partir duquel on définit les modèles d'utilisateurs. Un modèle de domaine est un graphe des concepts abordés dans l'hypermédia, caractérisant l'ensemble des connaissances accessibles dans l'hypermédia. On utilise alors des modèles d'utilisateurs du type *Overlay user models*, qui partagent la même représentation que le modèle de domaine. Ce sont des vecteurs d'attributs (un pour chaque concept) qui représentent une mesure d'intérêt ou de connaissance de l'utilisateur dans les concepts (Da Silva 1998, De Bra 2003, Kavcic 2000). Les modèles sont mis à jour à partir de la navigation de l'utilisateur. On peut également faire de l'inférence dans ces modèles en utilisant le formalisme des réseaux bayésiens (Da Silva 1998, Henze 1999).

Dans ce papier, nous nous intéressons à la définition automatique d'un modèle de domaine pour un hypermédia quelconque, à partir de son contenu. Il s'agit d'une étape préliminaire pour construire des versions adaptatives de systèmes hypermédias quelconques. Nous présentons ici une approche qui permet d'apprendre automatiquement une hiérarchie des concepts abordés dans un corpus de documents textuels, à partir du contenu de ceux-ci. Cette hiérarchie traduit une relation de spécialisation/généralisation entre les concepts. Comme nous le montrons, cette approche fournit une vue hiérarchique alternative du site web. Cette représentation du contenu thématique d'un site web permet de définir des modèles utilisateurs plus appropriés, en utilisant par exemple le formalisme des modèles bayésiens pour l'inférence. En outre, cette représentation associée à un outil de visualisation (tel que les *Treemaps*) permet la maintenance et/ou l'analyse des modèles d'utilisateurs.

2 Découverte de concepts à partir d'une collection de pages

Nous étudions ici comment apprendre automatiquement une hiérarchie de concepts à partir d'une collection de documents (les pages d'un site web) selon une relation de spécialisation/généralisation. Plusieurs approches ont été développées en recherche documentaire pour la génération de hiérarchies, reposant souvent sur des techniques de classification automatique. Ces techniques utilisent les similarités entre documents, mesurées par exemple sur des représentations vectorielles de fréquences de termes. Ce type de hiérarchie a été utilisé pour l'aide à la navigation et à la recherche d'information. Ces méthodes regroupent donc les documents en se basant uniquement sur leur similarité. A chaque niveau de la hiérarchie correspond des regroupements de plus en plus importants fusionnant les groupes de niveau inférieur suivant leurs similarités. Ce type de hiérarchies ne peut pas aisément être utilisé pour inférer des relations sémantiques nommées entre les concepts identifiés car il n'existe pas de relations sémantiques entre les nœuds des différents niveaux. Ces méthodes sont donc de peu d'utilité pour notre but. Récemment, de nouveaux types de hiérarchies construites automatiquement ont été proposés. Ce sont des hiérarchies portant sur les termes qui apparaissent dans les documents. Elles sont bâties à partir de relations de généralisation/spécialisation entre termes découvertes automatiquement sur le corpus exploité (Krishna 2001, Lawrie 2001, Sanderson 1999). 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. Nous proposons une 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, ces concepts reflètent mieux les sujets abordés dans la collection et offrent une description plus riche que des termes simples.

Nous décrivons maintenant notre méthode. Pour la clarté de la présentation, nous considérons qu'un hypermédia est composé d'unités (e.g. pages d'un site web) que nous appellerons des documents. L'approche consiste en une succession d'étapes. L'ensemble des documents est d'abord prétraité et segmenté en paragraphes homogènes. Cette tâche consiste à identifier dans chaque document des segments homogènes ou des frontières correspondant à des changements de thématique (Hearst 1997, Klavans 1998). Ces segments sont ensuite regroupés pour former des groupes de segments liés à une même thématique. Chaque thème découvert est alors considéré comme un concept de la collection. Un sous-produit de cette étape est que chaque thème est représenté par ensemble de mots clés. Les documents peuvent alors être classés par rapport aux concepts qu'ils abordent. On identifie alors des relations de spécialisation/généralisation entre concepts en utilisant une mesure de subsomption. Dans tous les traitements, documents et paragraphes sont représentés classiquement par des vecteurs de fréquences de leurs termes *tf-idf*, et la similarité entre deux entités (documents ou paragraphes) est calculée par le cosinus entre les vecteurs les représentant. Nous revenons maintenant avec un peu plus de détails sur chacune de ces étapes.

Il existe une littérature conséquente sur la segmentation car ce type de module est utilisé dans de nombreuses applications de recherche d'information (Klavans 1998). Nous considérons ici la segmentation des documents en passages cohérents et homogènes (Hearst 1997) afin d'identifier les thèmes dans un document. Nous avons employé pour cela une technique de segmentation thématique proposée dans (Salton 1996). 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. La méthode consiste à calculer les similarités entre les différents paragraphes du document et à ne retenir que celles qui sont supérieures à un certain seuil, puis à construire un graphe de similarités et à en extraire des triangles. Un tel triangle lie trois paragraphes et est susceptible de représenter une thématique cohérente. Ensuite, on fusionne itérativement les triangles dont la similarité est supérieure à un seuil. Une fois que chaque document est décomposé en un ensemble de thèmes, on procède à un regroupement des thèmes pour retenir un ensemble minimal de thèmes couvrant le corpus. Chaque thème est représenté par un ensemble de mots clés, les mots les plus fréquents dans le thème. Dans la suite, nous identifierons la notion de concepts aux thématiques découvertes de cette façon et représentées par des mots clés.

L'idée principale de notre méthode est d'apprendre automatiquement des relations de généralisation/spécialisation entre les concepts identifiés. Récemment, (Sanderson 1999) a proposé une méthode pour induire de telles hiérarchies entre les termes, basée sur la subsomption entre termes. La subsomption tente de mettre en avant les caractéristiques des différents termes et leurs relations. La subsomption caractérise une relation de généralité/spécificité entre deux termes et est basée sur un principe de co-occurrence non symétrique : Un terme x subsume un terme y (ou x est plus général que y) $\Leftrightarrow P(x/y) > t$ et $P(y/x) < P(x/y)$, où t est un seuil fixé.

Nous avons étendu la mesure de subsomption de termes ci-dessus à la subsomption entre thèmes. Cela nécessite le calcul de probabilités conditionnelles $P(C_i/C_j)$, la probabilité qu'un document traitant du concept C_j traite du concept C_i . Une fois les relations de spécialisation/généralisation détectées entre paires de concepts, nous appliquons la transitivité pour établir la hiérarchie de concept. Pour décider qu'un document D traite ou pas le concept C , on doit estimer $P(C/D)$ la probabilité du concept C sachant le document D ce qui n'est pas immédiat. Nous proposons de procéder à l'estimation des $P(C/d)$ via un algorithme EM qui calcule itérativement les probabilités $P(t/C)$ pour tous les concepts C et les termes du vocabulaire t , à travers la maximisation de la vraisemblance des documents de la collection. En faisant l'hypothèse d'un modèle naïve Bayes pour les documents, cela permet de calculer $P(d/C)$ et ensuite $P(C/d)$ via la règle de Bayes. Avec cette définition de la subsomption, un concept peut avoir plusieurs parents. Les différents chemins d'accès à un concept correspondent aux différents sens du concept et reflètent donc sa polysémie.

3 Application à la découverte de modèle de domaine

Nous avons appliqué notre méthode à la découverte d'un modèle de domaine, c.-à-d. une hiérarchie de concepts, sur une collection de documents. Les données utilisées pour cette expérience sont une sous hiérarchie du site www.looksmart.com, constituée d'environ 100 documents et 7000 termes sur l'intelligence artificielle. L'intérêt principal de ce corpus est que nous pouvons comparer, après l'apprentissage, la hiérarchie découverte et la hiérarchie manuelle d'origine. Nous donnons dans la table 1 des exemples de concepts extraits, identifiés par des ensembles de mots-clés. Comparée à la hiérarchie originelle de Looksmart en cinq catégories, la hiérarchie dérivée par notre algorithme est plus grande et plus profonde, la plupart des catégories originales sont raffinées. Par exemple, beaucoup de sous-catégories émergent de la catégorie originale « représentation de la connaissance » : ontologies, construction d'ontologies, etc.

Apprentissage d'une hiérarchie de concepts

1	definition AI intelligence learn knowledge solve build models brain Turing Test
2	informal formal ontology catalog types statements natural language names axiom definition logic
3	FCA techniques pattern relational database data mining ontology lattice categorie

TAB. 1 - Exemples de concepts (ensemble de mots-clés) extraits du corpus Looksmart.

Une caractéristique intéressante de cette organisation hiérarchique autour des concepts est qu'elle permet l'utilisation d'outils de visualisation efficaces. Nous considérons ici l'utilisation de *Treemaps* introduits dans (Schneiderman 1992), qui permettent de représenter une structure d'arbre dans un espace 2D. Chaque noeud est représenté par un rectangle dont la taille ou la couleur est déterminée par une valeur spécifique au nœud (ici un concept). La figure 1 montre une *Treemap* représentant le modèle de domaine du corpus Looksmart. Chaque concept est un rectangle, la hiérarchie est représentée par l'inclusion des rectangles. Une partie des mots-clés associés aux concepts est également affiché.

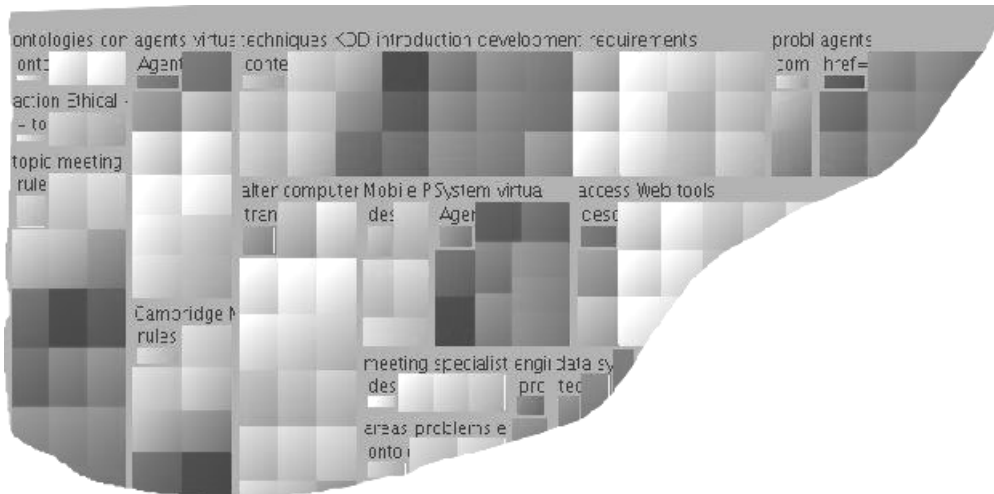


FIG. 1 - Partie de la *Treemap* de la hiérarchie de concepts découverte sur le corpus LookSmart. Des mots-clés sont affichés pour les concepts les plus généraux.

4 Utilisation pour la modélisation utilisateur

Comme illustré ci-dessus, notre méthode permet de définir automatiquement un modèle de domaine à partir des documents d'une collection. Elle peut alors être utilisée pour construire des modèles de domaine pour des hypermédia ou des sites web quelconques. Une fois qu'un modèle de domaine est appris, un modèle utilisateur du type *overlay user model* peut être défini, partageant la même représentation que le modèle de domaine. Des techniques standards incluant les réseaux bayésiens peuvent alors être utilisées pour apprendre et mettre à jour ces modèles utilisateurs (Da Silva 1998, Henze 1999), il s'agit d'une perspective de ce travail. Nous avons réalisé une expérience en appliquant notre méthode sur la collection des pages du site web d'un musée français en cours d'élaboration. La figure 2 montre le modèle de domaine résultant sous forme d'une *treemap*. Pour illustrer

comment un modèle utilisateur de ce type pourrait être utilisé, nous avons dessiné une session de navigation d'un utilisateur particulier sur cette *treemap*, où la couleur d'un concept (un rectangle) est fonction de la similarité thématique du concept avec les concepts des trois dernières pages visitées par l'utilisateur. Comme on peut le voir, les concepts qui sont près des pages récemment visitées par l'utilisateur sont proches du concept courant. D'autres informations pourraient être visualisées, les *Treemaps* permettent en effet de redéfinir facilement la couleur et la taille de rectangles. On pourrait donc analyser rapidement les centres d'intérêts d'un utilisateur en étudiant sa *Treemap* dans laquelle la couleur d'un rectangle (concept) est liée à son intérêt pour ce concept. On pourrait donc parcourir et étudier un modèle utilisateur en affectant une information de *connaissance* ou une information d'*intérêt* à la taille ou à la couleur des rectangles. Ce genre de visualisation permet d'avoir des informations globales et synthétiques sur un utilisateur.

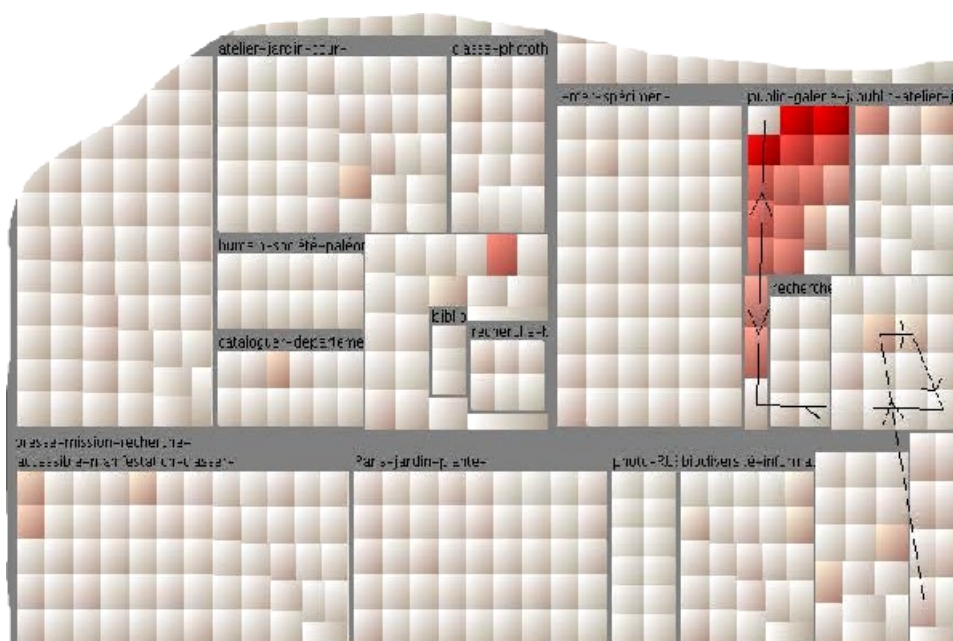


FIG. 2 - Partie de la *Treemap* du site web d'un musée français sur laquelle est plaquée une session de navigation. La couleur des concepts est liée à la proximité thématique des concepts avec les dernières pages visitées par l'utilisateur.

5 Conclusion

Nous avons présenté une approche pour apprendre automatiquement un modèle de domaine à partir d'un corpus de documents. Ce modèle est constitué d'un ensemble de concepts organisés en hiérarchie suivant une relation de spécialisation/généralisation. Nous avons illustré la méthode proposée sur deux corpus hypermédia. A partir de tels modèles de domaine, on peut définir des modèles utilisateur du type *overlay models* consistant en un ensemble d'attributs (e.g. intérêt, connaissance) associés aux concepts du domaine. Ces modèles utilisateurs peuvent être appris et mis à jour à partir de logs de navigation des

Apprentissage d'une hiérarchie de concepts

utilisateurs en utilisant par exemple le formalisme des réseaux bayésiens. Nous avons également montré comment des techniques efficaces de visualisation telles que les *Treemaps* peuvent être employées pour visualiser les modèles de domaine et utilisateur.

Remerciement: Les auteurs remercient Jean-Daniel Fekete (LRI, Université Paris Sud, Orsay) pour son aide et ses conseils sur les outils de visualisation et les *Treemaps*.

Références

- Brusilovsky P. (1996), Adaptive Hypermedia, an attempt to analyse and generalize, In Multimedia, Hypermedia, and Virtual Reality, Lecture Notes in Computer Science.
- Da Silva P., Van Durm V., Duwal E., Olivie H., (1998), Concepts and documents for adaptive educational hypermedia: a model and a prototype, Workshop on Adaptive Hypertext and Hypermedia.
- De Bra P., Aerts A., Berden B., De Lange B., Rousseau B., (2003), Aha! The adaptive hypermedia architecture, HyperText, United Kingdom.
- Hearst M., (1997), TextTitling: Segmenting Text into multi-paragraph Subtopic Passages. Computational Linguistics, pp. 33-64.
- Henze N., Nedjl W., (1999), Student modeling in an active learning environment using bayesian networks, User Modeling.
- Kavcic A., (2000), The role of user models in adaptive hypermedia systems, Electrotechnical Conference, MELECON.
- Klavans J., McKeown K. R., Kan M. Y., (1998), Ressources for Evaluation of Summarization Techniques. First International Conference on Language Ressources & Evaluation (LREC), Espagne, pp. 899-902.
- Krishna K., Krishnapuram R., (2001), A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining, International Conference on Information and Knowledge Management, pp.571-573.
- Lawrie D., Croft B., Rosenberg A., (2001), Finding Topic Words for Hierarchical Summarization, 24th annual international ACM SIGIR conference.
- Salton G., Singhal A., Buckley C., Mitra M., (1996), Automatic Text Decomposition Using Text Segments and Text Themes, Hypertext, pp. 53-65
- Sanderson M., Croft B., (1999), Deriving concept hierarchies from text. SIGIR Conference '99, pp.206-213.
- Schneiderman B., (1992), Tree visualization with tree-maps: 2-d space-filling approach, ACM Transactions on Graphics, Vol. 11, n° 1.

Summary

We describe a method for learning a concept hierarchy from a corpus. Concepts are identified by sets of keywords and hierarchically organized via a specialization/generalization relation. Such a hierarchy may be used to build automatically a domain model for various hypermedia and websites. We propose some ideas about how to define and learn user models based on such domain models. User and domain models may be visualized and analysed using visualization tools such as *Treemaps*. We provide experimental results on two corpus.