Automatic learning of domain model for personalized hypermedia applications

Hermine Njike, Thierry Artières, Patrick Gallinari, Julien Blanchard, Guillaume Letellier

LIP6, Université Paris 6
8 rue du capitaine Scott, 75015, Paris, France
{Firstname.Lastname}@lip6.fr

Abstract

This paper deals with the automatic building of personalized hypermedia. We build upon ideas developed for educational hypermedia; the definition of a domain model and the use of overlay user models. Since much work has been done on learning user models and adapting hypermedia based on such models, we tackle the core problem: the automatic definition of a domain model for a static hypermedia.

1 Introduction

In adaptive hypermedia one aims at developing user centric help strategies. One has to characterize the user in order to infer relevant help action for him all along his situation. A domain model may be used that characterizes the whole knowledge accessible in the hypermedia; it is used to infer information in the user model [De Bra et al, 2003]. Based on a user model, an adaptation model personalizes the hypermedia in order to offer to the user the most interesting and relevant information [Brusilovsky 2001]. An interesting task in adaptive hypermedia concerns educational hypermedia [Henze et al, 1999]. Although building such systems is still difficult, the task is well identified; in such systems domain models are often manually designed and defined as a graph of the concepts being discussed in the hypermedia [Herder et al, 2002]. Popular user models are *overlay* user models; these share the same representation as domain models and are used to represent a user knowledge / interest in a concept space. These models are vectors of attributes (measures of interest or knowledge), one for each concept in the domain model. These are updated from user navigation logs according to the domain model with standard techniques such as Bayesian Networks. However, the definition of domain models, hence user models, is done manually. We are interested in learning automatically such models, allowing the automatic building of personalized hypermedia from its content. This would allow for instance the automatic building of an adaptive website from any document collection. To do this, one can take advantage of works done in the educational hypermedia field concerning the learning of overlay user models and the personalization of hypermedia based on overlay user models. In this context, the core problem for the automatic conception of a personalized hypermedia lies in the automatic learning of a relevant domain model. We focus on this task here. This is not an easy task since this model involves high level concepts that cannot be easily inferred automatically.

2 Discovering concepts from a corpus

Some approaches were developed in the information retrieval field for generating documents hierarchies according to specialization/ generalization relations between concepts but these fail to infer semantic relations between concepts and are thus useless regarding our goal. Recently hierarchies have been proposed [Krishna et al, 2001, Sanderson et al, 1999] which are automatically built from corpora. They are term hierarchies built from generalization / specialization relations automatically discovered between terms in a corpus. Once this hierarchy is built, documents may be "projected" on it, thus producing a document hierarchy. We propose to extend these approaches to the construction of concept hierarchies where concepts are discovered from the corpus and are identified by sets of keywords and not only by single terms. Such a richer representation allows discovering relations that single term concept cannot. Our method takes as input a corpus, i.e. a collection of documents (e.g. pages of a website) that are preprocessed as usual in information retrieval tasks; non informative words are removed, all remaining words are lemmatized. The main steps of the procedure are as follows. Documents are first segmented into homogeneous paragraphs. The segmentation task consists in identifying, in each document, homogeneous text regions [Hearst, 1997]. Next, we cluster all these parts of documents in order to determine groups of paragraphs related to a similar topic. Each cluster is then considered to be a set of part of texts dealing with a concept of the collection. Each concept is represented as the set of most significant words (e.g. with highest *tfidf* values) in the texts of the cluster. Finally, we discover generalization links between concepts using a subsumption measure that we have extended to concepts. At last, applying transitivity leads to a concept hierarchy.

3 Application to user modeling

We applied our method to the discovery of a domain model, i.e. a concept hierarchy, on a collection of documents which is a part of the www.looksmart.com site hierarchies. The main interest of this corpus is that we can compare, after learning, the discovered hierarchy and the manually designed one. The corpus consists in about 100 documents and 7k terms about artificial intelligence. It has been manually organized in hierarchies of themes. We ran the method on the flat corpus, without any use of the hierarchical information. Compared to the initial Looksmart hierarchy with five categories, our discovered hierarchy is larger and deeper. Most of the original categories are refined. For example, many sub-categories do emerge from the original “Knowledge Representation” category: ontologies, building ontologies, KDD (papers about data representation for KDD)… It is clear that such a hierarchy could not have been obtained using single keyword concepts. An interesting feature of this discovered hierarchical organization is that it allows using efficient visualization tools such as Treemaps [Schneiderman, 1992]. Treemaps allow displaying a tree-like structure in a 2D grid where each node is represented by a rectangle whose size or color determined by a node specific value. Fig. 1 shows a Treemap representing the discovered Looksmart domain model. Each concept is shown as a rectangle with different colour; the hierarchy is shown through inclusion of rectangles. Smaller rectangles represent documents. Set of keywords associated to concepts only are shown for readability reasons.

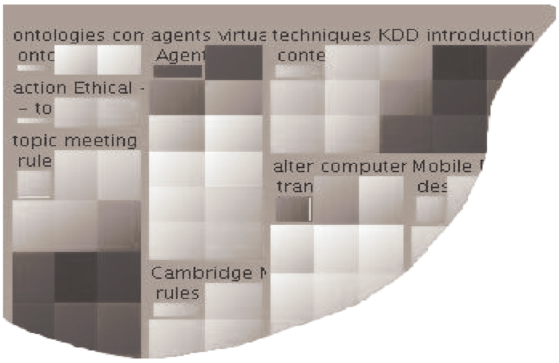

Fig. 1. Part of the treemap of the concept hierarchy discovered on the Looksmart corpus. Each rectangle stands for a concept.

We realized a second experiment by running the method on the pages of the website of a French museum. Fig. 2 shows the resulting domain model with french keywords. As an illustration, we have shown a navigation path of a particular user on this treemap, where the colour of a concept is a function of its thematic similarity with the three last pages visited by the user. As may be seen, concepts close to these pages lie close to the current concept. Other information could be visualized. Hence, one can browse and investigate a user model by assigning for instance *knowledge* or *interest* information about concepts to the size or colour of rectangles.

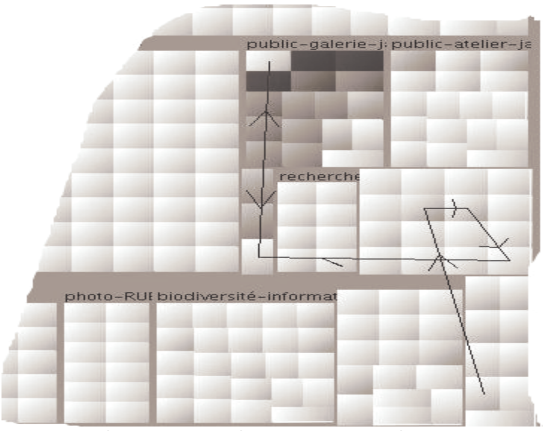

Fig. 2. Part of the treemap for the website of a French museum where a navigation path has been drawn.

4 Conclusion

We described an approach to automatically learn a domain model from a corpus of hypermedia documents. This model consists of concepts organized into a specialization / generalization Based on such a domain model, one can define user models as overlay models, and use existing techniques for learning and personalization. Our procedure may be thought as a core step for developing automatically an adaptive hypermedia from any document collection.

Acknowledgments

The authors would like to thank J.D. Fekete from Université Paris Sud (France) for helpful discussion about treemaps.

References

[Brusilovsky, 2001] Brusilovsky P., Adaptive Hypermedia, User Modeling and User-Adapted Interaction, 2001.

[De Bra et al., 2003] De Bra P., Aerts A., Berden B., De Lange B., Rousseau B., Aha! The adaptive hypermedia architecture, HT’03, United Kingdom.

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

[Henze et al., 1999] Henze N., Nedjl W., Student modeling in an active learning environment using bayesian networks, User Modeling 1999.

[Herder et al., 2002] Herder E., Van Dijlk B., Personalized adaptation to device characteristics, International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 2002.

[Krishna et al., 2001] Krishna K., Krishnapuram R., A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining. International Conference on Information and Knowledge Management, 2001, Atlanta, Georgia, USA. pp.571-573.

[Sanderson et al., 1999] Sanderson M., Croft B., 1999, Deriving concept hierarchies from text. SIGIR Conference '99. pp.206-213.

[Schneiderman, 1992] Schneiderman B., Tree visualization with tree-maps: 2-d space-filling approach, ACM Transactions on Graphics, Vol. 11, No. 1, January 1992.