

Enabling Ontology Switching in Browsing Interfaces

Thomas Mandl, Christa Womser-Hacker

University of Hildesheim, Information Science, Marienburger Platz 22
D-31141 Hildesheim, Germany
{mandl, womser}@uni-hildesheim.de

Abstract. User interfaces with hierarchical browsing structures have proved to be easy to use. They are usually based on an ontology which organizes the terms of the domain. In order to be able to use such an interface the user should be able to adapt the hierarchical structure to his knowledge, his mental model of the domain and his vocabulary. For that reason, the user should be able to choose his preferred knowledge structure by switching the hierarchical ontologies. Only the use of the appropriate terminology guarantees access for all user groups. This paper discusses the advantages of the approach and the necessary underlying transfer relations between the different vocabularies. It gives a practical example of a vague relation which has been established by machine learning algorithms between two library catalogues for information science. These transfer relations can be exploited for the implementation of a virtual library shelf where the distribution of books over the shelves is not fixed but flexible. A change of the hierarchical system results in an automatic re-organization of the material.

1 Introduction

Software should be adapted to the needs of users. This includes the selection of the appropriate terminology for each user. Only a significant conceptual overlap between the users and the systems terminology guarantees highly usable information systems. Personal preferences may be based on education, occupation, computer experience, age and culture.

Browsing is an important strategy in information retrieval [12]. It is especially suited for the non expert user because it does not require the production of keywords but solely the recognition of promising entries to be followed. The disadvantage of browsing interfaces is the occupation of large amounts of screen space. For interfaces for experts this is a serious drawback. However, in ubiquitous access, browsing plays and will play a major role. Since browsing is an informal and sometimes unpredictable approach, it requires careful design of the access methodology and interface. The main design issue in browsing systems is the semantic organization of the knowledge objects and the navigation within them.

Not all users share the same conceptualization of objects in a domain. People assign different meanings to the same word quite commonly. The effectiveness of browsing interfaces is effected by semantic heterogeneity. The integration of various collections of documents is a main advantage of digital libraries. Large virtual collections can be created. However, most original databases are based on their specific content representation. Often, they are associated with an ontology optimized for their content. The organization of these ontologies represents a view of the domain.

Often, novel ontologies are developed for virtual libraries. A new ontology needs to be learned by all users. Their knowledge about other ontologies cannot be activated in order to improve their interaction with a virtual digital library including several collections. Ontology merging as discussed in [13] is a typical solution which creates a novel system as well.

Semantic heterogeneity needs to be handled more effectively. Several solutions have been developed for keyword search in heterogeneous collections. The heterogeneity is resolved by creating relations between entries in the ontologies. The two main strategies are intellectual assignment and machine learning systems based on training data. The latter results in a higher level of vagueness which does not necessarily lead to a worse performance.

For the user interface, two levels of vagueness for semantic relationships between categories need to be considered. The hard and unique assignment in list oriented systems is relaxed by associative maps.

2 Semantic Heterogeneity of Ontologies

The structure and organization of an ontology is the manifestation of a view on the world or the domain. A considerable number of ontologies exists in every domain and their number is growing. Many ordering systems have concepts in common, however, because they arose within a certain context in response to specific demands. The same term may have a significantly different meaning in other domains [1], [2], [8].

Although this situation is quite natural and not likely to change, it complicates or even prevents the successful communication between communities using different ontologies. The same is true for the exchange of documents between different groups within one digital library. This especially applies for different countries and cultures. The world-wide exchange of documents involves ontologies from different cultures which are sometimes organized extremely different. Strict models for standardization are hardly ever a feasible solution, because the heterogeneity is a natural cause of different viewpoints.

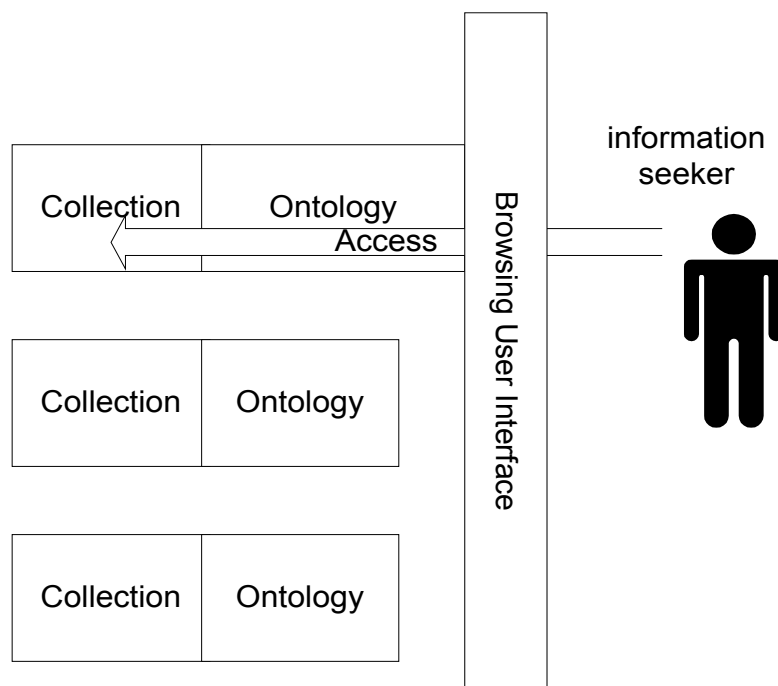


Fig. 1. Separation of ontologies results in separation of collections outside virtual libraries

3 Dealing with Semantic Heterogeneity

Each user wants to be able to access as many knowledge objects as possible with the same interface and the same terminology. Furthermore, we claim, he should be able to choose the most appropriate terminology for himself and be able to access all objects with the selected vocabulary. Since the knowledge objects are usually represented within the context of their database by its specific ontology or also by its specific free text vocabulary, not all objects are represented by the same ontology. Even when the same terms occur, their meaning may be different in different contexts.

To deal with this semantic heterogeneity between ontologies, semantic transformation modules between different document description languages are necessary. The mapping between different terminologies can be done by relying on intellectual or machine derived transfer modules.

Intellectual transfers use cross-concordances between different classification schemes or thesauri. They allow a fairly safe and reliable mapping between terms from different documentation languages. Intellectual cross-concordances are considered as a precise transfer relation. They allow different types of relations to be combined and they are usually associated with a higher quality than machine generated relations. Their major disadvantage is the high cost. Not all institutions participating in distributed digital libraries can afford them, or are willing to commit the necessary resources. Therefore, other – cheaper – ways had to be found to create transfer relations. They are usually derived by statistical or machine learning algorithms which exploit multiply indexed objects. Therefore, an overlap of objects between the involved ontologies is necessary. In contrast to the intellectual transfer rules, the relations derived by machine learning are vague and give degrees of conformance between terms.

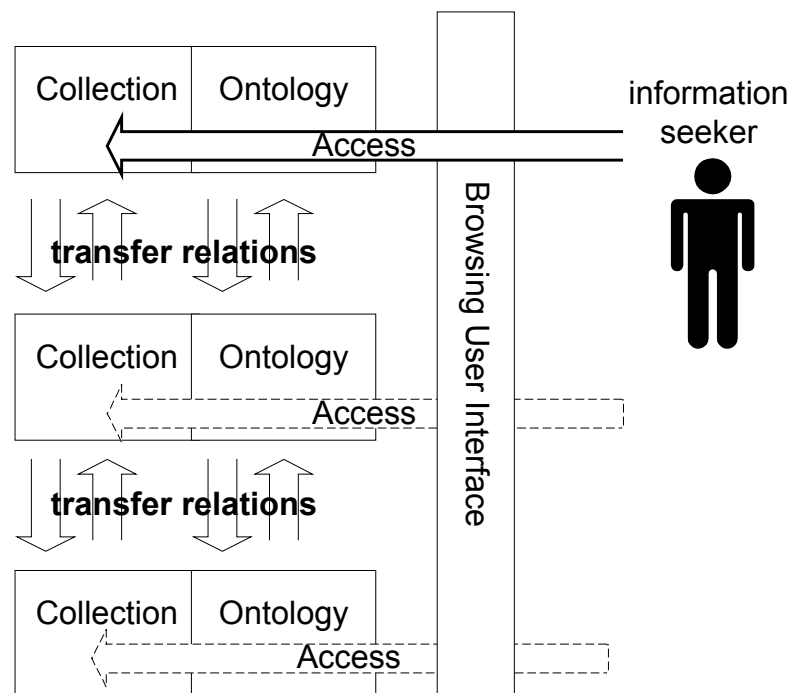


Fig. 2. MyShelf enables access through one ontology only. Collections represented by other ontologies are connected through transfer relations

Often such intellectual transfer relations are available because e.g. library personnel sees the necessity to support users who need to work with different ontologies. However, these systems are rarely available electronically and if, they are not fully linked in order to reach the documents and to support browsing in different systems.

Limitation to these certain and relatively secure methods would lead to a situation in which transfers between most ontologies would be impossible. Furthermore, vague transformation relations include the following advantages:

- They exploit human information work which has already been carried out.
- Intellectual indexing has been carried out and should not be neglected by replacing it with full-text indexing only.
- In cases where intellectual information work has been done repeatedly by indexing a document in several institutions with different ontologies, these multiple knowledge structures support the transfer relations.
- They allow associative connections which may have not been sanctioned by a domain expert but which still may reflect a user need in a certain situation.

3.1 Machine Learning

Machine learning methods offer a general, automatic way to create transfer relations on the basis of bibliographic data. E.g. statistical transfer modules can be used to supplement or replace cross-concordances. They allow a crosswalk between two different thesauri or even between a thesaurus and the terms of automatically indexed documents. Learning algorithms exploit the analysis of co-occurrence of terms within two sets of comparable documents.

Machine learning in text categorization research provides an appropriate technology for this task. Mostly, text categorization assign documents to predefined categories based on a full text analysis [14]. Texts are indexed with standard information retrieval methods and represented by weights assigned to words or terms based on their frequency of occurrence. These terms can be regarded as features. Co-occurrence analysis is the basis of those methods. It takes advantage of the fact that the content analysis from two different libraries for a document held in both collections will represent the same semantic content in different ontologies. The terms from content analysis system A that occur together with terms from content analysis system B can be extracted. The assumption is that the terms from A have a similar semantic as the related entries from B, and thus a vague transfer relation is established.

Neural networks [4] showed to be a successful method to learn mappings for transfer modules in heterogeneous digital libraries [9]. Their performance can be improved by combining them with other algorithms.

3.2 Applications of Learning Methods to Transfer Modules

Different learning methods have been applied to form transfer modules between ontologies. Most often, statistical association measures like Naive Bayes map between pairs of terms. These learning algorithms derive the knowledge from examples provided as training data and do not rely on further human contributions. An overview is given in [6]. For the treatment of semantic heterogeneity with machine learning, methods from text categorization are an important starting point [7].

Crestani and Rijsbergen present a backpropagation network for a mapping between different queries in which the representation schemes are equivalent. The same architecture can be modified for general transformations between heterogeneous representation schemes. A query was transformed into queries which achieved better results. Improved queries for training were found using relevance feedback. The transformed queries achieved similar retrieval quality compared to the original query [3]. The process can be seen as a query extension.

Both interface types can be implemented at different levels. However, the traditional browsing interface based on hard assignments is more appropriate for very realistic applications which try to copy traditional library buildings.

5 The virtual library shelf for information science

The virtual library shelf MyShelf enhances the browsing access to heterogeneously represented objects. It is applied to the information science books in the library of the University of Hildesheim. MyShelf integrates the relevant library stock, other libraries and teaching material in the internet. The access is possible through various hierarchical ontologies and especially library catalogues. The ordering system can be chosen by the user and the system reorganizes its content accordingly. For that purpose, transfer relations between the objects and categories need to be established.

5.1 Information Science at the University of Hildesheim

Information Science has only recently been introduced at the University of Hildesheim. It is a central subject within the curriculum International Information Management [11]. As a consequence, Information Science is not listed as a discipline in the library catalogue. Relevant books are placed under the signatures of several other disciplines like economics, computer science, linguistics, psychology, sociology, design and mathematics. Therefore, the books are located in different shelves in the library and it is difficult to get an overview over the books available. Informal interviews with students showed that this leads to difficulties for the access of library material. This situation is typical for situations in which universal access for heterogeneous user groups with different backgrounds is envisioned.

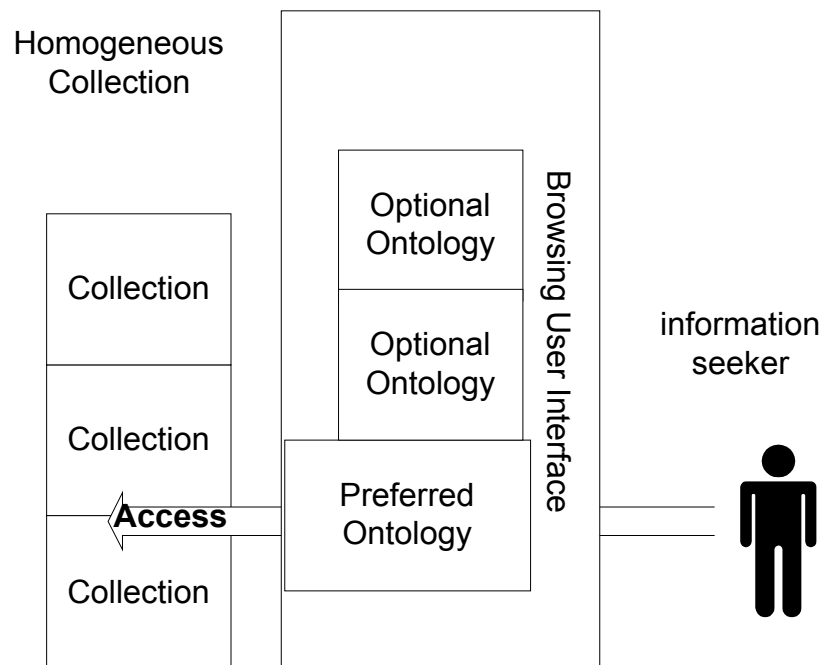


Fig. 4. The users' perception of MyShelf as a homogeneous system. Neither the heterogeneity of the collections nor the heterogeneity of the ontologies needs to be regarded

5.2 The MyShelf-Model

As a solution, a virtual library shelf based on the MyShelf model is being developed. It integrates several ontologies for information science. Since the system is virtual, it will not be limited to the literature available at the local university library. For a later phase, the integration of teaching material provided by faculty is envisioned. This material is intensely used by students and its importance is increasing. The location within a larger pool of literature at a certain position in an ontology will allow students to browse to related documents and get acquainted with the organization structure of the ontology.

Optionally, the user can browse the local or both local and remote collections. The disadvantages of a necessarily limited supply are overcome. Students may navigate through the collections of both university libraries using their favorite ontology of information science. Each user can switch the ontology at any point and can still access the data from both libraries.

5.3. Transfer Relations

In order to identify the books relevant for information science, several heuristics have been applied [5]. A large list of books possibly relevant was determined by searching all catalogue categories containing terms from information science. These terms were extracted from the library catalogues of the universities in Konstanz and Saarbrücken, where information science is recognized as a discipline. The large corpus was compared to a complete list of books from the library of Konstanz. As a result, some 6000 relevant books were found. These had a notation from the catalogue in Konstanz assigned to them as well as a notation from Hildesheim. Therefore, this set was well suited for the automatic construction of transfer relations between the two libraries.

However, a closer look revealed several challenges. The set represented 141 terms from Konstanz and 481 from Hildesheim. Considering a transfer from Hildesheim to Konstanz we found out that on average we had some 35 training examples for each class. This would be a satisfying value for a machine learning algorithm. However, the standard deviation is rather high and reaches 102. In other words, there are many classes with very few examples. In the whole set 19% of all categories of the Konstanz catalogue are represented only by one example and 28% by one or two examples. In addition, there are many categories which are not present in the data at all. On the other side, there are some categories with very many training examples. This will result in a high bias for these categories in training.

As a result, a machine learning algorithm will not achieve good results for many categories. This situation is typical for real world data which has not been set up for such tasks. For example, many of the categories of the Library of Congress Subject headings have never been even used.

As a consequence, the choice on whether to use vague transfer relations or intellectually derived relations cannot be decided for a whole collection. Instead, we propose a hybrid transfer scheme. Categories with good results from machine learning approaches can be handled by automatically derived relations. On the other hand, classes with either too many or too few examples need to be further evaluated. In case they are not very relevant because they do involve many objects, they can be neglected. If they are used as descriptors for many objects they need to be subjected to an intellectual transfer.

6 Resume

This paper described the problems arising in the context of semantic heterogeneity in browsing user interfaces. In order to assure universal access for many users, different viewpoints need to be integrated in information systems. Transfer relationships established upon examples or intellectually derived form the basis for ontology switching interfaces. A case study for the establishment of transfer relations for information science library catalogues is presented.

References

1. Buckland, M.; Gey, F. et al. 1999. Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine* 5 (1) <http://www.dlib.org/dlib/january99/buckland/01buckland.html>
2. Chen, H. 1998. Introduction: Trailblazing Path to Semantic Interoperability. *Journal of the American Society for Information Science. JASIS* 49 (7). pp. 579-581.
3. Crestani, F.; Rijsbergen, K. van 1997. A Model for Adaptive Information Retrieval. *Journal of Intelligent Information Systems* 8 (1). pp. 29-56. <http://www.cs.strath.ac.uk/~fabioc/papers/97-joiis.pdf>
4. Ham, F.; Kostanic, I. 2001. *Principles of Neurocomputing for Science & Engineering*. Boston et al.: McGraw-Hill.
5. Hanke, P.; Mandl, T.; Womser-Hacker, C. 2002: Entwurf eines Virtuellen Bibliotheksregals für die Informationswissenschaft. *Proceedings 8. Intl. Symposium für Informationswissenschaft. (ISI 2002)*. Oct. 2002, Regensburg, Germany.
6. Hellweg, H.; Krause, J.; Mandl, T.; Marx, J.; Müller, M.; Mutschke, P.; Strötgen, R. 2001. *Treatment of Semantic Heterogeneity in Information Retrieval*. Technical Report, IZ Sozialwissenschaften, Bonn. http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab23
7. Joachims, T. 1998: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *European Conference on Machine Learning (ECML)*. pp. 137-142.
8. Krause, J.; Mandl, T.; Stempfhuber, M. 1997. *Text-Fakten-Integration in ELVIRA*. Technical report, IZ Sozialwissenschaften, Bonn. <http://www.bonn.iz-soz.de/publications/series/working-papers/index.htm#Text>
9. Mandl, T. 2000. Tolerant Information Retrieval with Backpropagation Networks. *Neural Computing & Applications*. 9 (4). pp. 280-289.
10. Mandl, T; Eibl, M. 2001. Evaluating Visualizations: A Method for Comparing 2D Maps. In: Smith, M; Salvendy, G; Harris, D; Koubek, R (eds.): *Usability Design and Interface Evaluation: Cognitive Engineering, Intelligent Agents and Virtual Reality. Proceedings of the HCI International 2001 (9th International Conference on Human-Computer Interaction)*, New Orleans, August 2001. Mahwah, NJ; London: Lawrence Erlbaum Associates. Vol. 1. S. 1145-1149. Mandl Eibl HCI 2001
11. Mandl, T; Womser-Hacker, C. 2001. Currículos de Ciência da Informação na Alemanha. In *Tecnologia da Informação e a Questão Social. XXI Congresso da Sociedade Brasileira de Computação (SBC 2001) IX Workshop Sobre Educação em Computação (WEI)*.
12. Marchionini, G., 1995. *Information Seeking in Electronic Environments*. Cambridge.
13. Noy N; Musen, M. 1999. SMART: Automated Support for Ontology Merging and Alignment. *Twelfth Workshop on Knowledge Acquisition, Modeling and Management*. Banff, Alberta, Canada. 16-21 October, 1999. <http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Fridman1/NoyMusen.pdf>
14. Yang, Y., Liu, X. 1999. A re-examination of text categorization methods. *22nd Intl ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 42-49.