A Graphical Interface for Document Retrieval Based on Formal Concept Analysis

Uta E. Priss
School of Library and Information Science, Indiana University
Bloomington, IN 47405. Phone: 812 855 2793. Email: upriss@indiana.edu

June 13, 2002

Abstract

The research presented in this paper is part of a larger project of building a graphical interface for document retrieval which is based on a conceptual analysis of explicit and implicit structures among document data. Mathematical lattices, which are formalized according to Formal Concept Analysis, serve as an internal model. The lattice model is opposed to the more traditional vector space-based information retrieval models. Besides the document database a thesaurus is included in the system and also formalized as a lattice. The system obtains a high degree of flexibility, i.e. adaptability to different document databases or knowledge bases and to changing users preferences, from its modular design and from a set of lattice combination options.

1 Introduction

The background for the research presented in this paper is an attempt to build a graphical interface for document retrieval. Lin (1997) presents a detailed survey on how visualizations can enhance information retrieval by letting users browse through a graphical representation of the requested documents. Thus a query is resolved interactively between the system which graphically displays implicit and explicit structures among the document data and the users who use their knowledge to make decisions which the system cannot make itself. The system we propose to build can be distinguished from existing systems with regard to two major points: first, instead of the more common vector-space model we use a lattice-based model for the document data. We cannot provide many details in this paper about the differences between the two models, but it should be sufficient to say that, for visualization purposes, lattices can be displayed in two-dimensional space without loss of information whereas vector spaces of higher dimensions, which must be dimensionally reduced or projected before they can be visualized in the plane, usually cause loss of information. Second, the difference between our model and the lattice-based modelings of Carpineto & Romano (1995) and Godin et al. (1993) is that we enhance the retrieval process by including a knowledge base or thesaurus.

Information retrieval processes can involve different degrees of user interaction. In one extreme, the user sends a query to the system and the system returns a result. If users are unsatisfied with the result they must formulate a new query and send it back to the system. The problems of this approach are usually difficulties in the query formulation and the linear display of the query result (compare Lin (1997)). In the other extreme, users formulate queries interactively with the system. For example, the system provides starting points which the user selects and modifies. Following the user's selection the system provides further query choices for the user until the query is completed. Search results can be provided at the completion of the query or during each stage of query formulation. For example, the numbers of documents that are relevant at each stage of a query can be displayed so that users know whether they have to broaden or narrow their search. The system therefore provides query aids preferably in the form of a graphical display which assists the user in identifying relevant documents. For this approach it is sufficient for the system to retrieve documents with high recall and low precision, since the decision about the relevance of documents is achieved by the user.

The interface we propose in this paper represents a hybrid system that combines a certain 'intelligence' of the system with user decisions instead of having the system making autonomous decisions. As explained in the preceding paragraph, this approach seems to be one possibility for overcoming the precision/recall bottleneck of information retrieval systems. The system's intelligence is expressed in its ability to recognize and graphically represent implicit relationships among the documents (or other data). In the simplest version, only information about documents and their keywords, which is easily accessible, is used. More advanced versions can incorporate natural language processing components to provide a deeper analysis of the document descriptions. In the modeling which we propose in this paper, a thesaurus or knowledge base assists in the semantic conceptual analysis. Software agents can be applied to tune the system to the user so that user feedback is incorporated into the formulation of new queries. The selected information can then be filtered according to preferences which the user requests from the system.

The graphical interface for information retrieval described in this paper uses Formal Concept Analysis (Ganter & Wille, 1996) for its formal modeling. Formal Concept Analysis provides graphical representations in the form of mathematical lattices that are generated from the analysis of formal objects and their formal attributes. Carpineto & Romano (1995) and Godin et al. (1993) have built independently of each other similar interfaces using Formal Concept Analysis, but they do not incorporate a knowledge representation or classification system into their systems. They claim that lattice-based retrieval using only the information presented by documents and their keywords is already superior to boolean retrieval. Since traditional manual information retrieval always incorporates the use of classification systems, thesauri, or library catalogues it seems that by using a knowledge base or thesaurus the retrieval software can simulate some of the common sense reasoning involved in manual information retrieval and thus combine the advantages of manual and computerized retrieval. Some computerized information retrieval systems, such as the Internet search engines 'Yahoo!' and 'Infoseek', already use classification systems for their retrieval processes.

The system that we are planning to build consists of several modular components. The thesaurus component and the document database component are maintained inde-

pendently of each other so that they can be exchanged depending on the application areas. Only the keyword (or class) mapping component which maps document keywords to classes of the thesaurus has to be modified if a new thesaurus or document database is used. This may be done semi-automatically by using a synonym dictionary. The formal structures of the system and the software do not have to be changed. In this paper we formally define the main components of the system. Several details, such as browsing and searching features, graphical display algorithms, and so on are left for future publications. And the evaluation of the system's performance cannot be discussed at this point since the system is not implemented yet.

2 A Document Retrieval System

The document retrieval system based on Formal Concept Analysis which we propose for the graphical interface consists of several components. The document database is represented as a formal context[1] $\mathcal{K}_D := (D, C_D \cup A_D, I_D)$ where D is the set of documents, C_D is a set of keywords of the documents, A_D is a set of further attributes of the documents, such as "year of publication" and "number of pages", and I_D is a relation. An example of a small document database is presented in Figure 1a. The formal attributes can be divided into the sets $C_D := \{$catalogs, classification, Internet, hardware$\}$ and $A_D := \{$published in$\}$ where 'published in' is a so called 'many-valued' attribute. The other attributes are single valued (binary). The concept lattice in Figure 1b corresponds to a context (D, C_D, I_D). It displays how the books and keywords are related. To find all books which contain a certain keyword, every path from the keyword to the bottom of the lattice has to be followed. A book which appears lower in the lattice than another book contains all keywords of the higher book (compare 'book 2' and 'book 3'). If a user is looking for books on 'hardware' and discovers that there are too many books, he or she can check the subconcepts of the node which denotes 'hardware' to see which nodes have lesser amounts of books attached, but refer to other useful keywords. The effectiveness of a manual keyword search often depends on the user's knowledge of the semantic relations among the keywords. For example, since users know that the World Wide Web is part of the Internet, they conclude that Internet-related books may contain information on the WorldWideWeb and vice versa. These relationships may appear in the document lattice, for example, if the keywords 'Internet' and 'World Wide Web' are shared by several books, however for a consistent display of implicit semantic information it is usually better to combine the document lattice with a thesaurus lattice (see below).

[1] Compare Ganter & Wille (1996) for the definition of a formal context. For this paper it is enough to state that a formal context is denoted by $\mathcal{K} := (G, M, I)$ and represented by a cross table of the set G of formal objects, the set M of formal attributes and a relation I between objects and attributes. An equivalent representation of a formal context is a concept lattice $\underline{\mathcal{B}}(\mathcal{K})$.

	catalogs	classific.	hardware	Internet	published in
book1	x	x			1995
book2			x		1980
book3			x	x	1995
book4				x	1996

Figure 1a: A document database context

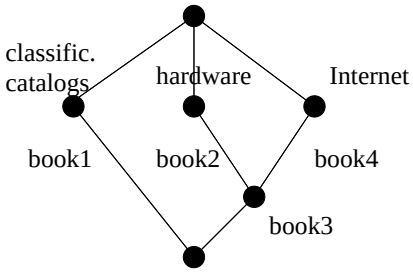

Figure 1b: A document database concept lattice

The concept lattice in Figure 1b is the kind of lattice Carpineto & Romano use in their paper. They do not include lattices for combined sets of attributes, such as $C_D \cup A_D$, or many-valued attributes, such as 'published in'. In our system combined sets of attributes and many-valued attributes are included so that users can, for example, query for all documents on a certain subject written at a certain time, written by a certain author, or having a certain minimal number of pages. Furthermore, the different combinations of attribute sets allow for the system to be flexible concerning the users' interests.

The second component of the system is the thesaurus or knowledge base which is formally defined as a formal context $\mathcal{K}_T := (C_T, C_T, I_T)$. The classes of the thesaurus are taken as formal objects and formal attributes and C_T denotes the set of classes. The concept lattice is constructed as the Dedekind closure of the classes ordered by the subclass-superclass relation (compare Priss (1996)). An example for a formal context and concept lattice of a thesaurus is represented in Figures 2a and 2b. A mapping $M : C_D \longrightarrow C_T$ assigns classes of the thesaurus to the keywords of the documents. In the example of Figures 1 and 2, the identical mapping is used.

	comp. sci.	inform. sci.	library sci.
catalogs			x
classific.		x	x
hardware	x		
Internet	x	x	

Figure 2a: A thesaurus context

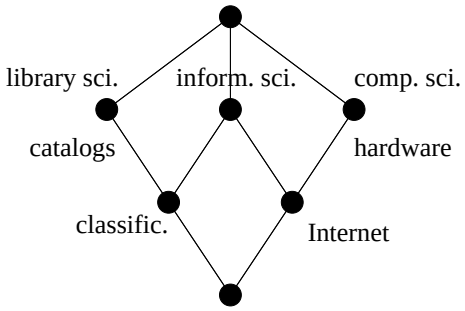

Figure 2b: A thesaurus concept lattice

There are several ways to construct concept lattices from a document database and a thesaurus. The first possibility is to map the documents to the hierarchy of the thesaurus. Formally this is done by constructing the formal context $\mathcal{K}_{T/D} := (C_T \cup (D, C_D), C_T, I_T)$ (compare Figure 3a). (D, C_D) denotes the pairs consisting of documents and their keywords. The relation between the pairs in (D, C_D) and C_T is defined as follows: $(d, c_d) I_T c :\Longleftrightarrow c_d I_T c$. The lattice of $\mathcal{K}_{T/D}$ is equivalent to the lattice of $\mathcal{K}_T$, but has the pairs consisting of documents and their keywords as further formal objects. It is often sufficient to display the numbers of documents for each node (compare Figure 3b) and allow the users to click on the document numbers if they want to obtain more details (for example titles) of the documents.

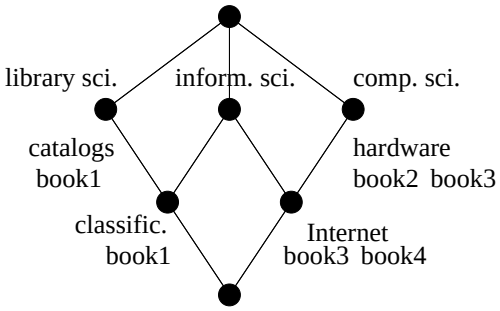

Figure 3a

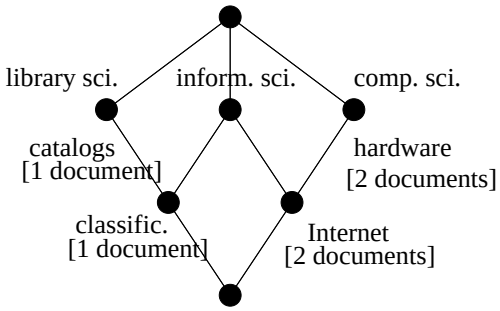

Figure 3b

Although the lattice in Figure 3b displays how many documents belong to each keyword, it does not show which documents share their keywords. The relationships can be investigated by composing the document lattice with the thesaurus lattice (Figures 4a and 4b) or by other constructions (Figures 5a and 5b). The context in Figure 4a consists of the identity context in the upper left corner, the document context (D, C_D, C_D) in the lower left corner, the thesaurus context (C_D, C_T, I_T) in the upper right corner, and the context composition $(D, C_T, I_D \circ I_T)$ in the lower right corner. Figure 4b shows the concept lattice for the context in Figure 4a.

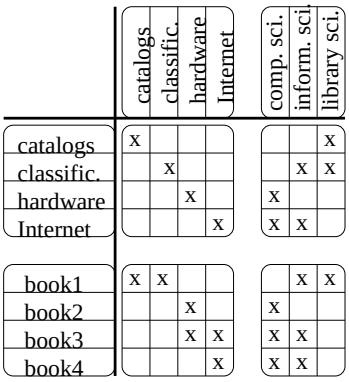

Figure 4a

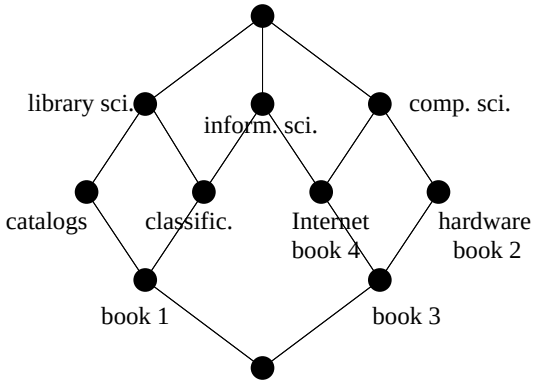

Figure 4b

Documents can also be viewed using more general classes than the classes to which their keywords match. Figures 5a and 5b show the documents from Figure 1 in relation to the high level classes 'library science', 'information science', and 'computer science' according to the thesaurus in Figure 2. Viewing more general classes involves a quantification on the object level as defined in Relational Concept Analysis (compare Priss (1996)). A decision has to be made as to whether a document belongs to a high level class if all its keywords belong to that class (the context $(D, C_T, \overline{I_D \circ \overline{I_T}})$, Figure 5a), if most of its keywords belong to the class, or if at least one keyword belongs to that class (the context $(D, C_T, I_D \circ I_T)$ Figure 5b).

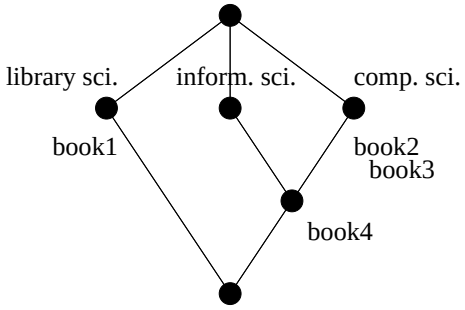

Figure 5a

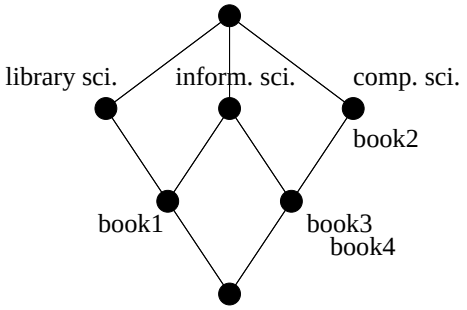

Figure 5b

A further option for the user is to combine several concept lattices into 'nested line diagrams' (Ganter & Wille, 1996). Figure 6a shows an example in which the many-valued attribute 'published in' is scaled into 'recent' and 'old' books. The outer structure contains the documents and their keywords. The inner structure shows the age of the books. In a larger application users would have a choice between several scales. A software called TOSCANA (Vogt & Wille, 1994) manages the combination of the scales. Up to four scales can be combined in any order. By clicking on nodes the users can display or hide the more detailed information contained in the inner structures of the nested line diagram. This option can be used to filter information according to user preferences.

	catalogs	classific.	hardware	Internet	recent	old
book1	x	x			x	
book2			x			x
book3			x	x	x	
book4				x	x	

Figure 6a: A scaled version of the context in Figure 1a

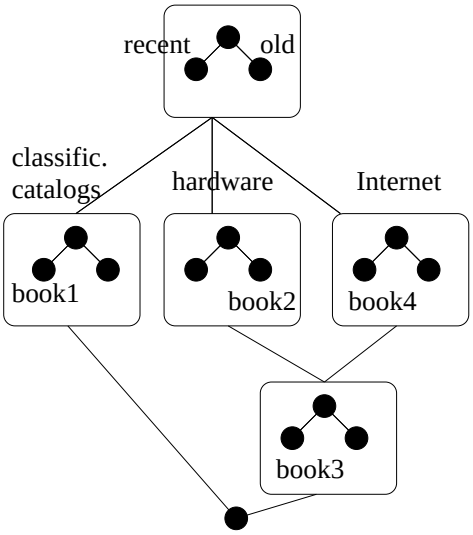

Figure 6b: A nested line-diagram of the context in Figure 6a

3 Conclusion

The examples in this paper obviously represent 'toy examples'. Their lattices are small enough to be completely displayed. In large scale applications it is usually not possible to display complete lattices of document databases or thesauri. There are several possibilities to obtain partial lattices for display purposes, for example, 'neighborhood lattices' (Priss & Wille, in prep.) can be computed. They consist of a subset of the formal objects, their attributes, and the other formal objects of their attributes. Other options to reduce lattices include selecting smaller sets of attributes, the 'bounding method' which Carpineto & Romano (1995) propose or fisheye views (Furnas, 1986).

To summarize, lattice-based retrieval models allow a graphical display of relations among documents and their keywords. By applying thesauri to the retrieval process, common sense semantic knowledge can be coded into the system. In this paper we describe a retrieval system which serves for interactive query formulation or browsing. Future research has to investigate how measures or probabilistic models can be combined with the lattices so that some of the search algorithms, which have been developed for vector-space retrieval, can be implemented for lattices. To decide whether lattice models are fully equivalent or superior to vector space models not only concerning interactive searching, but also concerning fully automated searching is also left for future research.

4 References

Carpineto, Claudio; Romano, Giovanni 1995. ULYSSES: A Lattice-based Multiple Interaction Strategy Retrieval Interface. In Blumenthal et al. Human-Computer Inter-

action, Springer Verlag.

Furnas, G. W. 1986. Generalized fisheye views. *Proc. of the Human Factors in Computing Systems*:16-23.

Ganter, Bernhard; Wille, Rudolf 1996. Formale Begriffsanalyse: Mathematische Grundlagen. Springer-Verlag.

Godin, R.; Missaoui, R.; April, A. 1993. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. *Int. J. Man-Machine Studies* 38:747-767.

Lin, Xia 1997. Map Displays for Information Retrieval. *J. of the Amer. Soc. of Inf. Sci.* 48:40-54.

Priss, Uta E. 1996. Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases. Dissertation, TH-Darmstadt.

Priss, Uta; Wille, Rudolf, in prep. A Formalization of Roget's International Thesaurus.

Vogt, Frank; Wille, Rudolf 1994. TOSCANA - A Graphical Tool for Analyzing and Exploring Data. TH-Darmstadt, Preprint 1670.