Metadata for Multidimensional Categorization and Navigation Support on Multimedia Documents

Thomas Klement, Matthias Hemmje

German National Research Center for Information Technology (GMD)
Integrated Publication and Information Systems Institute (IPSI)
[klement, hemmje]@darmstadt.gmd.de

Abstract

An increasing technological effort is spent on integrated representations of document collections and metadata. For instance the emerging XML standard offers opportunities to represent metadata in for, e.g., improving query and navigation support within web-based document collections. Despite this development, most applications of catalogue metaphors on the web ranging from small web site catalogues to complex search and browsing engines have not yet significantly changed their underlying data model.

In information systems applications, the "catalogue" concept is mostly used as a connotation for information systems that categorize content by means of document annotation hierarchies. Within user interfaces, catalogue metaphors are in wide spread use. To take advantage of more flexible and efficient catalogue representations based on multidimensional data models a project about an integrated Information Catalogue Environment (ICE) is currently performed at GMD-IPSI. The goal of the ICE project is to provide efficient and flexible catalogue oriented categorization and representation mechanisms for multimedia document types and structures.

This goal is achieved by modeling categorization metadata of a multidimensional nature to support data exploration by means of multidimensional catalogue navigation.

1 Introduction

The amount of multimedia documents served by the internet is almost exponentially rising from year to year. Finding and accessing information served by this fast growing information infrastructure has become more comfortable over the last years. However, nowadays information technologies still offer only rather simple mechanisms of categorizing documents to enable navigation and browsing support which is often a suitable alternative to full text search approaches. The importance of such navigational exploration mechanisms supported by underlying catalogue categorization and representation structures is motivated by two examples.

The LookSmart ([LookSmart]) search engine is providing category-based navigation services on the World Wide Web and evolved very successfully in the recent past. This categorical navigation engine has been integrated with many of the current web search engine products (like Digital's AltaVista, @Home, CompuServe, Netscape, Microsoft, Bigfoot, Desktop News, HotBot). This emphasizes the need for providing hierarchical navigation support structures while exploring WWW content. Looksmart's cascading menu interface provides users with fast, intuitive access to relevant online content. Furthermore, looksmart hosts the world's largest editorially reviewed database of web content with 350,000 site listings in 20,000 categories.

Categorical navigation engines like Looksmart can be considered as "pull" technologies because users trigger each navigation or data transfer to the client side explicitly. A popular contrary example is the channel or "push" technology (Netscape, Microsoft). In this case the server controls the data transfer to the client. The user subscribes to categorized channels that have a similar structure related to "pulling" catalogue systems.

The main drawbacks of current applications of categorization techniques for large catalogues of documents are that the category hierarchies are **deeply nested** and categories usually have **navigation-context dependencies** within their tree hierarchy**.**

An example shows that these two problems are tightly coupled. At the time this paper was written, the Looksmart catalogue contained 4 different entries of the category "news" below the 12 main categories, 37 "new" entries on the third level and so on, which points out a redundancy problem. The category "news" was defined heterogeneously on different levels. Obviously the provider of the document catalogue had to split up the categories in this highly redundant way to decrease the number of redundant occurrences of categorization terms as well as the overall number of levels and terms per level. Such problems result from the limitation to a one-dimensional categorization hierarchiy of the document collection. Despite of wide spread

demands for categorical exploration support, many state-of-the-art user interfaces lack advanced categorical navigation techniques that reduce or completely overcome the problems outlined above.

The next chapter describes how a multidimensional modeling approach can help to reduce such problems and how the resulting requirements towards the underlying implementation in a database systems are related to multidimensional OLAP database technologies.

2 Inherent Benefits of a Datacube Model for Catalogue Applications

The metadata schema for multidimensional categorization and navigation support that is introduced in the following is based on a so called "datacube" data model (datacube models are often referred to as "hypercubes", too). In parallel to introducing the basic concepts of datacube models some general advantages of its application are outlined. These primary benefits are the basis for the advantages of many common OLAP applications, except of the significant difference that analysis functions are less important for catalogue environments than navigation, selection and viewing functions. ([Codd], [Thomson])

The fundamental elements of a datacube are its so called "dimensions" which are structured hierarchically. In general a dimension hierarchy consists of an arbitrary formed tree of nodes (so called "members" or "positions"). Members in the context of catalogue applications are informational categories. Hence a dimension is a tree structure of informational categories.

The basic goal for multidimensional modeling of hierarchies is to reduce the total number of members, and as a result to minimize the number of categories the user has to navigate for content exploration. This means, that from the users´point of view, they are supported to navigate through a number of categorical document annoation hierarchies in parallel. Following this approach, a potentially huge number of documents can be addressed within the space spanned by possible combinations of positions in each of the dimensions. To achieve this "spanning" property, the dimension hierarchies have to be modeled independent of each other to prevent redundant member semantics. So members with a “certain” semantic in one dimension should not be part of other dimensions with a different semantic.

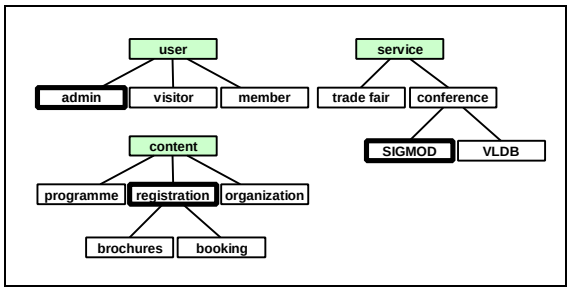

Figure 2-1

The example in figure 2-1 describes the dimensions of a trade fair & congress organizer's catalogue datacube. The members in bold boxes in figure 2-1 can be reached by navigating to a "multidimensional address" (MDA) that references documents which are positioned (or categorized) within each defined dimension. In the trade fair & congress organizer’s catalogue hypercube the MDA {admin, registration, SIGMOD} might reference documents that list registered participants of, e.g., the SIGMOD conference. Thus a multidimensional address has the following syntax:

MDA = { dim_0,dim_1 ... dim_n }

for n defined dimensions and dim_k ($0 \le k \le n$) are members of dimension k. More generally, a collection of dimensions creates a space that is called a datacube.

Fehler! Keine gültige Verknüpfung. Figure 2-2

The datacube itself consist of single **cells** that are referenced by MDAs. Such a datacube applied to an exemplar *trade fair & congress organizer's catalogue application* is visualized in figure 2-2. It is created by three dimensions for *users*, *services* and *content* as allready mentioned in figure 2-1. Note that not only **leaf members** in a dimension create a row or column in the datacube and its corresponding cells but that also **inner members** of a dimension are reflected in corresponding cells as well. This property of the datacube model enables the positioning of informationally more general documents (like, e.g., introductions) on higher level MDAs in multidimensional catalogue information systems.

If the expressive properties of the datacube model are analized w.r.t. supporting the general requirements

(a) intuitive design of the user interface

(b) context sensitive navigation within the overall document space

(c) fast document access

(d) easy adaption of the user interface from information systems of different application areas

of catalogue applications, the datacube model appears as a datamodel that is capeable of supporting the outlined application requirements quite well.

Obviously, the main drawback (according to the requirements (a) and (c)) of introducing a multidimensional catalogue metaphor to naive users is that this metaphor is unknown to them. However, as the number of financial and business applications in the area of OLAP constantly increases at a significant rate, multidimensional variants of allready well established user interface metaphors will gain more and more ground, too. Taking this motivation as a starting point, applications of multidimensional categorization within catalogue metaphors have advantages worth to be considered:

Advantages w.r.t. requirement (a):

The categorization of documents into multiple dimensions is straightforward because the semantic of members within a dimsnion is unique. For that reason, the automatic or administrative categorization and the corresponding positioning of documents within the datacube is less costly. The following chapter introduces the description of a template mechanism supporting a flexible and efficient positioning of documents in the datacube.

Advantages w.r.t. requirement (b):

The semantic of all members within dimensions is precisly defined, because the relationship of members to their ancestor members within a multidimensional catalogue is defined as a “part of” relation. A "part of relation" is considered to be an informationally homogenious structuring method, whereas relationships to ancestors in onedimensional catalogue applications are mostly of mixed types and therefore considered as informationally heterogenious. The result is a better orientation of the user within the overall document space.

Advantages w.r.t. requirement (c):

The average duration of naive user's catalogue navigation and document access task has not yet been investigated in formal experiments w.r.t. onedimensional approaches. However, there are two reasons why the multidimensional navigation and access method is expected to be more efficient: The selection of relevant members in the navigated dimensions is done straight forward, because the semantic of each member is unique. Moreover, the number of drill down and roll up operations needed for multiple navigation and access tasks is reduced if the navigation contexts (resulting MDA) of successive tasks are similar.

Chapter 4 introduces a novel approach for the adaptation of navigation support mechanisms for multiple dimensions to the degree of user's experience w.r.t. interacting with multidimensional catalogues. The use of so called "dynamic menus" allows the design of a variety of apllication specific user interfaces for navigation considering user groups with experiences ranging from naive users to expert users.

The content of Chapter 5 outlines the fundamental importance of precalculated metadata aggregations stored as attributes of datacube cells. Metadata aggregates, for instance the number of available documents within a datacube cell, are not only important for fast generation of simple navigation support mechanisms (e.g., navigation menus) but also for advanced navigation support mechanisms (e.g. .early navigation feedback, i.e., qualitative or quantitative "look ahead" functionalities)

3 Document Positioning and Specialization

A representation of a datacube cell can (according to chapter 2) be defined by a pair of (MDA, document) values. Obviously, it would be appropriate to enable the mapping of documents to a specific volume of cells in the datacube (a so called "subcube" of the entire datacube) if they belong to more than a single cell. Consider the example of figure 3-1 where a document is mapped into the cells of the MDA volume {admin, content, trade fair},{admin, programme, trade fair} and {admin, registration, traid fair}.

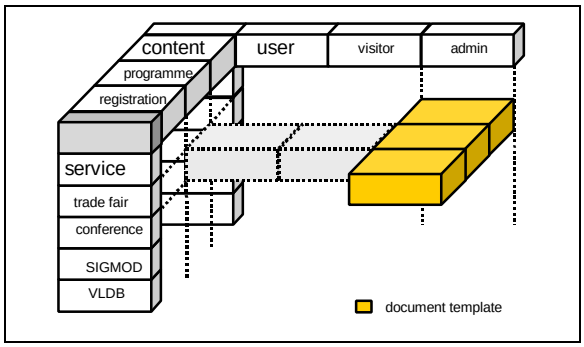

Figure 3-1

In the light of the well known OLAP experience multidimnesional applications have to deal with sparse storage of data and under the assumption that a large extend of document cell mappings (MDA, document) can be substituted by document subcube mappings the ICE approach provides a **template positioning** mechanism that allows the definition of a mapping via (MDA pattern, document) value pairs.

The **MDA pattern** has the syntax

$MDA_p = \{ dim_0 [*], dim_1 [*] \ldots dim_n [*] \}$

for n defined dimensions and dim_k ($0 \leq k \leq n$) are members of dimension k. The stars, enclosed by brackets, reflect the optional use of a MDA pattern for the inclusion of all subtree members below the

member dim_k. Hence, MDAs are a subset of MDA patterns.

The valid MDA pattern that summarizes the three MDAs listed in the example above is {admin, content*, traid fair}.

In general a MDA pattern

$MDA_p = \{dim_0\ [*], dim_1\ [*] \ldots dim_n\ [*]\ \}$

is **valid** for a navigation position

$MDA_{current} = \{\ dimc_0, dimc_1 \ldots dimc_n\ \},$

if each $dimc_k$ $(0 \le k \le n)$ is equivalent to dimk or dim_k* specifies a subtree of members where dim_k is an anchestor of $dimc_k$

Within the ICE approach, the algorithm that processes MDA patterns for template positioning is kept very simple for performance reasons. In particular it has to fullfill the two conflicting goals of economic storage consumption for mapping pairs and of efficient access to datacube cells (i.e., the efficient execution of the MDA pattern match operation). However, it is powerful enough to ease the document positioning significantly.

When using "part off" relation types between members and their ancestors the ICE approach allows for an additional navigation feature called **document specialization**. The document specialization mechanism allows for a dynamic representation of different levels of informational specialisation.

Imagine that the two mappings ({admin, programme, traid fair}, document1).and ({admin, content*, traid fair}, document2). were inserted into the datacube. If users reach the navigation context specified by the MDA {admin, programme, traid fair} during an exploration, the first MDA pattern matches this position more specifically than the second pattern .

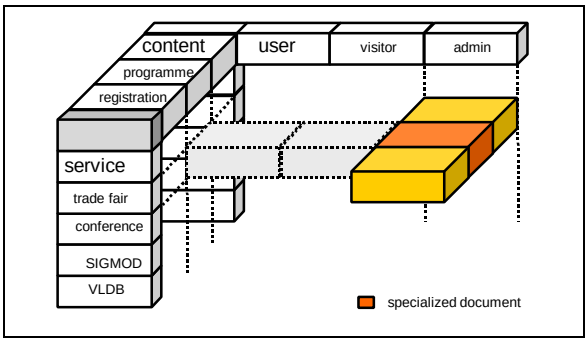

Figure 3-2

Thus users should obtain the two corresponding documents ranked by the order document1, document2 or just a single search result of document1.

The following definition describes a more precise priority rule which might be used to implement a rspecialization mechanism:

Given the two mappings

$(MDA_{p1}, document_1), (MDA_{p2}, document_2)$

and the MDA patterns MDA_{p1}, MDA_{p2}

$MDA_{p1} = \{dim1_0\ [*], dim1_1\ [*] \ldots dim1_n\ [*]\ \},$
$MDA_{p2} = \{\ dim2_0\ [*], dim2_1\ [*] \ldots dim2_n\ [*]\ \}$

that are valid for the current navigation context

$MDA_{current} = \{\ dimc_0, dimc_1 \ldots dimc_n\ \}$

$document_1$ is more specific than $document_2$ according to $MDA_{current}$, if each dim_{1k} $(0 \le k \le n)$ is an anchestors of or equivalent to dim_{2k}.

Having defined how documents are accessed from a multidimensional catalogue datacube, the next chapter focuses on the ICE system's user interface that provides users with navigation and access support to the document collection.

4 Catalogue Navigation

The catalogue navigation support provided by the ICE system's user interface is described by an exemplar application for *the trade fair & congress organizer's catalogue datacube.* The chosen user interface layout is derived from state-of-the-art web-based catalogue layouts ([Priestley] describes hierarchical navigation from an user interface designers view, but limited to onedimensional hierarchies. Meanwhile the document markup standards for the World Wide Web have been substantially improved by XML ([7]). XML related standardization affects catalogue interfaces by additional facilities that are mentioned later on).

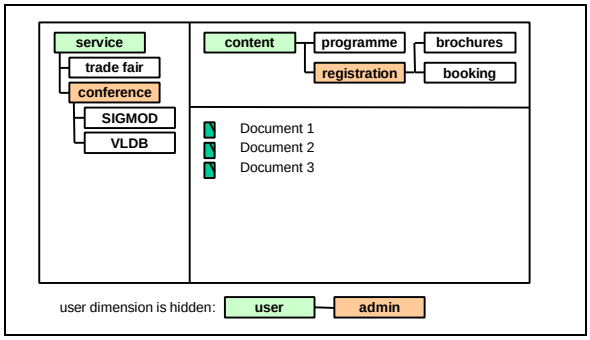

Figure 4-1

Figure 4-1 shows a navigation user interface for the dimensions *service* and *content* in *the trade fair & congress organizers catalogue datacube* example. The *user* dimension controls access permissions. It is currently hidden in the example user interface, because users are not allowed to change their permissions navigationally (Remark: In the ICE system, user authentication can be implemented by

separat authentication dialogs that set the appropriate member position in the user dimension implicitly).

The application window in figure 4.1 is divided into frames. The top window and the left window display **navigation menus** for the visible dimensions. Navigation menus can display dimension trees only partially, if parts of the tree are restricted to specific user groups. The navigation menus allow drill down and roll up operations while displaying the resulting dimension structures and navigation contetxts clearly. The colored menu members *conference* and *registration* indicate that the current MDA of the navigation process is {admin, registration, conference}. All members of a current MDA position are visualized including their next level submenu entries. The cell that is referenced by the current MDA in the trade faire organizer's datacube contains three documents that are visualized in the **document focus** frame (often called "content frame"). The document focus displays the exploration results to the user.

The example user interface design provides no **visual feedback** about the existence of accessable documents during the navigation **in advance**. If users select the member *SIGMOD* the focus could display an empty list of resulting documents, because the new addressed cell contains no documents. Obviously, there is a need for "navigation look ahead" feedback, because exploring subcubes in the document space, where no documents are available, doesn't make sence. To privide a more economically navigation, the ICE systems's catalogue interface components offer two basic types of navigation menus.

ICE's "static menus" visualize a navigation context of a current MDA. In static menus the members of the navigation context that refer to empty cells are visualized in a "navigationally deactivated" style in the resulting submenu. This approach prevents users from dealing with changing menu structures.

Figure 4-2 illustrates the side effects between static navigation menus at a selection of a menu entry.

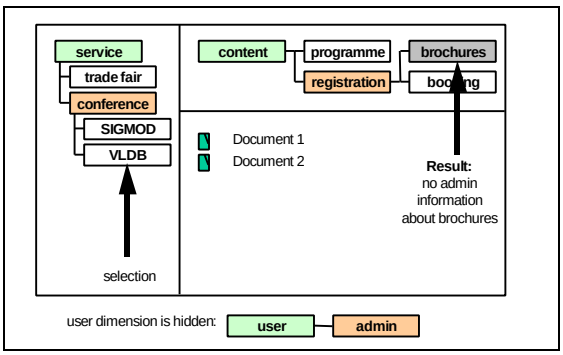

Figure 4-2

The current MDA is the same as in the example above. If the user selects the submenu item *VLDB* below *conference* the following MDA would be {admin, registration, VLDB}. As a result of the context switch, the submenu item *brochures* below registration is deactivated, because the MDA {admin, brochures, VLDB} references an empty cell of the datacube. The cell referenced by the drill down opportunity before (immediate selection of *brochures*) would have addressed a non-empty cell at the MDA {admin, conference, brochures}.

The characteristic of static menus is the property that every submenu entry is displayed, although it could be deactivated w.r.t. navigation. The usage of static menus is reasonable for dimensions that shall not change their structure for a better orientation in the corresponding navigation contexts. However, it is often more suitable to hide menu entries from the user to save screen space.

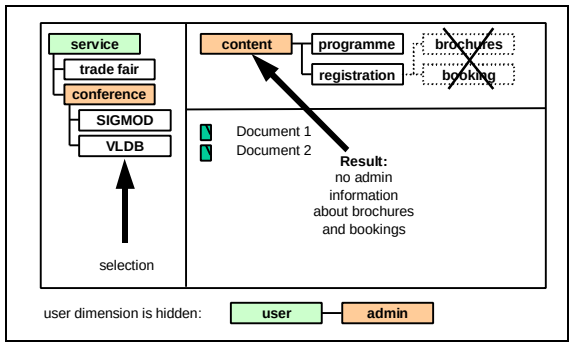

Figure 4-3

Therefore, ICE additionally offers "dynamic menus" that visualize submenus which refer only to non-empty cells in the datacube. This reqirement for dynamic menus causes a side effect described by figure 4-3. Again, the current MDA is {admin, registration, conference}, but at this time dynamic menus are used. The selection of *VLDB* has the effect that the resulting MDA is {admin, registration, VLDB}. The figure shows that any MDA obtained by a selection from the submenu of the menu item *registration* would reference an empty cell. So the *content* dimension has to be rolled up to obtain a valid state for the dynamic menu *content* after the *VLDB* drill down operation. This side effect ensures that all submenus address only non-empty cells in the datacube.

Summing up, a member of static menus is not effected by roll up/ drill down operations of other dimension menus, besides being displayed in an activated or deactivated style for further navigation. Dynamic menus may change their structure as a result of drill down and roll up operations in other dimensions. One can conclude, that dynamic menus visualize dimensions in a more efficient way (w.r.t. the use of screenspace) while they are less effective w.r.t. navigation support.. When performing a drill down operation on a certain dimension, users have to be aware of changing structures in other dynamic menus. To meet application requirements in a most flexible way, ICE allows the combined application of

static and dynamic menus within user interface designs. Furthermore, menu types can be changed at runtime to achieve a most flexible support for exploration activities.

5 Aggregates

The above described menus, implemented by the ICE system, ensure fast system response during navigation by means of aggregation pre-calculations.

The menu generation component would have to perform costly passes through member subtrees in certain dimensions, resolving some pattern matching tasks for checking the validity of mappings at the passed MDAs. This recursive calculation is to expensive for supporting highly interactive navigation in an efficient way, especially for large datacubes.

For example, the ICE system provides pre-calculated aggregations of document numbers for any MDA according to each defined dimension. The aggregated document numbers have to be calculated according to the following requirements:

In this exemplar case, an aggregate **A** has been pre-calculated for the dimension **d** (where the roll up side effect occurs as a result of a navigation in another dimension) and the MDA **m** (which will be the next MDA in the navigation context). This means, **A** is the number of documents that are referenced by a set **S** of MDAs "below" **m** with variable members for dimension **d** and fixed members for all other dimensions. The MDAs in **S** "below" **m** are composed from a member in dimension **d** that is the (side) effected member or a descendent of it and the fixed members of the remaining dimensions.

The calculation time for the aggregation values can be further reduced if aggregations are accumulated for higher level aggregation pre-calculations. Extensive research efforts have been spent on aggregation issues in the area of OLAP applications research. For example, [Harinarayan] focuses on the mentioned reuse of pre-aggregated values for building higher level aggregates.

Finally, the metadata of ICE do not only consist of data about dimension hierarchies for one or multiple menu structures. For efficiency reasons they furthermore contain aggregate values which are pre-calculated at the time a document reference is inserted into the catalogue's datacube. Such aggregation values can be explicitly visualized as a kind of "navigation look ahead" support. This can either be achieved by visualizing aggregtions as numeric values or menu entries can be visualized in a different style if the corresponding cell in the datacube doesn't refer to documents (because no document was inserted).

6 Future Work and Conclusions

ICE will offer additional aggregates in future, for instance probabilistic values that represent the relevance of the applied categorization term according to the referenced document content, average age of the documents, etc. Looking at currently ongoing standardization activities for metadata (like, e.g., [Dublin Core] and [RDF]) one can expect increasing support for automatic insertion of document references into a multidimensional catalogue. Taking advantage multidimensional catalogue representations, automatic or semi-automatic categorization mechanisms should become less difficult to implement, because of the independent semantics which are offered by multiple dimensions.

The ICE system will be utilized soon for improved access control in applications that are designed for the wide range of intranet, extranet and internet use in parallel. Access control in multidimensional information systems allows the assignment of privileges to a user dimension that models access permissions for different user groups in a natural multidimensional style. The privileges in such systems can be set according to an arbitrary number of characteristics (dimensions) which provides considerable advantages as well.

Since the multidimensional approach has been well-approved in OLAP applications over the last years, we expect that it will be beneficial for information catalogue applications as well.

7 References

[Codd] Codd, E. F., Salley "Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT- Mandate" Arbor Software

[Dublin Core] Dublin Core Metadata "http://purl.org/metadata/dublin_core/"

[Harinarayan] Harinarayan, Venky, Rajaraman, Anand, Ullman, Jeffrey "Implementing Data Cubes Efficiently" ACM SIGMOD 96, Montreal, Canada

[LookSmart] "http://www.looksmart.com"

[Orenstein] Orenstein, Jack A.
"Spacial Query Processing in an Object-Oriented Datebase System"
Computer Corporation of America, 1986

[Priestley] Priestley, Michael
"Navigation Issues in Hypertext: documenting complex hierarchies with HTML frames"
ACM SIGDOC 97, Snowbird Utah, USA

[RDF] Resource Description Framework
"http://www.w3c.org/RDF/"

[Thomson] Thomson, Eric
"OLAP Solutions: Building Multidimensional Information Systems",
Wiley Computer Publishing

[XML] Extensible Markup Language
"http://www.w3c.org/XML/"