Developing Digital Libraries Using

Data Warehousing and Data Mining Techniques

Cássia Blondet Baruque

Rubens Nascimento Melo

(e-mail: cassia, rubens@inf.puc-rio.br)

PUC-RioInf. MCC28/04 July, 2004

Abstract: We propose a manner to the development of Digital Libraries (DL), using Data Warehousing (DWing) and Data Mining (Dmining) Techniques. This DL will be a component of an E-learning environment and will assist the students in a specified course. This will help them in their studies and researches, once the Web will be already filtered by the Data Mining techniques on the subject they need to study. We propose components of the Data Warehousing architecture on which Data Mining techniques can be applied. The Data Warehouse will be like a Central Library and the Data Marts (portions of the Data Warehouse separated by a certain criteria - the course, for example) like Departament Libraries. The retrieval of the DL documents by students will be made through OLAP (On-Line Analytical Processing) techniques based on the DL catalog, which is modeled multidimensionally and based on the Dublin Core metadata elements. A discussion of the DL automatic refreshing based on information collected on the students' interactions is also presented.

Keywords: E-Learning, Digital Library, Data Warehouse, Data Mining, Learning Objects

Resumo: Nós propusemos uma maneira para o desenvolvimento de bibliotecas digitais (BD) utilizando Técnicas de Data Warehousing (DW) e Data Mining (DM). Esta Biblioteca Digital é um componente de um ambiente de e-Learning e servirá de apoio aos estudantes em um curso específico. Tal biblioteca os ajudará em seus estudos e pesquisas, uma vez que a Web já terá sido filtrada pelas técnicas de DM em relação ao assunto que eles precisam estudar. Nõs propusemos componentes da arquitetura DW nos quais técnicas de DM possam ser aplicadas. O DW será como uma Biblioteca Central e os Data Marts (porções do DW separadas por um certo critério – o curso, por exemplo) serão como Bibliotecas Departamentais. A recuperação de documentos da BD por estudantes será feita através de técnicas OLAP (On-Line Analytical Processing), baseadas no Catálogo da BD, que estará modelado multidimensionalmente e baseado nos elementos do Metadados Dublin Core. Também é apresentada discussão sobre o "refreshing" automático da BD, a partir de informações coletadas nas interações dos estudantes .

Palavras-chave: Aprendizado à Distância, Biblioteca Digital, Data Warehouse, Data Mining, Objetos de Aprendizado

Developing Digital Libraries Using Data Warehousing and Data Mining Techniques

1. Introduction

With the dissemination of the Internet, a great amount of documents is available for search and retrieval on the Web. According to [CHA95] the Internet is now one of the biggest information repositories. However, its content is disorganized and distributed. Moreover, the diverse hardware and software platforms, as well the different document formats and diverse media available compose a great heterogeneous database, which contains structured, half-structured and non-structured data. All this distribution and heterogeneity have contributed to make the search and the Web content acquisition difficult.

In this context, Digital Libraries (DL), a recent research area, aims at organizing and promoting an easier access to documents on the Web. As there are many definitions in the literature [Schwartz2001] and there is no consensus regarding the DL concept, in this work a DL is considered to be a great object collection, in diverse digital formats, persistent, managed and well organized using a catalogue and with access through the Web.

The development of a DL generally implies in integration of disbributed multimedia content on the Web. Since the hypermedia nature of the Web implies in navigation through the content in order to get the desired information, the organization of the integrated data should consider a content categorization into hierarchies.

There are various initiatives which aim at developing DL whose proposal is to solve the problem of content integration as well as the access to these contents using hierarchical classification [GAAS1999]. The following projects are some related works: Digital Stanford Library Technologies [PBCCG2000], Digital Illinois Library Initiative Project [Chen2000], Digital Alexandria Library Project [ACDFF+1995] and University of Digital Michigan Library Project [WB1998]. However, it is noted that such initiatives do not present a comprehensive proposal to address the issues related to DL.

A research area that has been contributing to solve complex database problems is the area of Data Warehousing (DWing). The DWing approach has been very useful to address issues related to data integration and complex search.

The process of digital library development includes issues such as the integration of complex documents found on the Web. Moreover, access to the DL must be assisted by the use of content hierarchies that guide the user in the discovery and filtering of information of his/her interest.

The study of new methods for the DL development is becoming a very fertile research area. In some research work more emphasis is being given to the item of integration of complex and heterogeneous data, using approaches such as CORBA, agents and mediators, as found in [PBCCG2000], [Chen2000], and [WB1998].

Some works consider the challenges related to the semantics of the DL [Chen2000] and others suggest the use of hierarchies classification for refinement of the users´s search, as mentioned in [GAAS1999].

Some works emphasize the need for a systematic approach that allows the automation of the main typical library functions, such as classification, cataloguing, etc. Moreover, the proposed solutions normally are based on proprietary data which can be catalogued using USMARC format that is typical of traditional libraries.

This work presents an architecture for DL development based on the Data Warehousing (DWing) approach and using some Data Mining (DMining) techniques. DWing has been used for data integration to give support to the decision-making process. As a result of the DWing, there is a database called Data Warehouse (DW), which is subject-oriented and promotes content organization into hierarchies in an integrated way. On the other hand, DMining allows to find, extract, filter and evaluate the desired information and digital objects, as well as tracking and analyzing the users standards´ accesses.

In some areas, such as the administrative, statistics and GIS (Systems of Geographic Information), the DWing approach has contributed to the integration of complex data and access to them in a satisfactory fashion. Similarly, this work contributes to the solution of some fundamental problems in this area.

In Section 2, the Data Warehousing (Dwing) is mapped to the "Data Librarying" (Dling) process. In Section 3 the Dling Components are presented. In Section 4 we propose the use of data mining techniques in some of the Dling Processes. Finally, in Section 5, some concluding remarks are made and we comment about the changes we made to the proposed work.

2. Data Warehousing (DWing) and Data Librarying (DLing)

2.1 - The DWing approach in the DL development

The DL development based on the Dwing approach implies in understanding the DWing architecture and how to use and/or adapt its processes and components for the DL.

2.1.1 Data Warehousing

In accordance with William H. Inmon [Inmon1996], a DW has the following characteristics:

◆ It is subject-oriented (the data are stored in accordance with specific areas of the business or specific subjects/aspects of the company interest).

◆ It is integrated (i.e., it integrates data from diverse sources, while identifying and correcting inconsistencies).

◆ It is a collection of non-volatile data (it means that data are loaded and accessed, but its updating does not occur in a DW environment).

◆ It is variant in the time (the time horizon for DW is significantly longer than that of production systems; the data consist of a sophisticated series of "snapshots" obtained at a certain moment; and the DW key structure always consists of some temporal elements).

◆ It is used for supporting management decisions.

Generally, an architecture for systems based on DW involves the integration of current and historical data. The data sources can be internal (operational systems of the company/institution) or external (containing complementary data originated from the organization, such as economic indicators). Generally, the data integration deals with different data models, definitions and/or platforms. This heterogeneity demands the existence of applications that extract and transform data in a way that the data integration becomes possible. Once integrated, the new data are stored in a new database - DW - that combines different points of view for supporting management decisions. This database is used for data analysis by final users. DW can be divided into some databases called Data Marts (DM) [Kimball96]. Such DMs contain information that can be useful to different departments of the company. They are also considered as departmental DWs. DM/DW can be accessed by OLAP (Online Analytical Processing) or DMining tools and/or DSS (Decision Support Systems). These tools make the data navigation possible, , as well as the managerial analysis and the knowledge discovery. An important component of this architecture is the metadata repository, where the information about the DW development can be found.

2.1.2 DLing from the DWing Approach

By applying the DW-related concepts shown above in the DL process, we observe that:

◆ a DL must be subject-oriented (as mentioned previously, it is important for the users that they can search for documents through a subject hierarchical classification).

◆ a DL must have an integrated view of documents. A possible distribution of these documents, as well as inconsistencies, must be transparent to the final user.

◆ Documents and its corresponding metadata must be loaded only one time in the DL and its contents do not have to be updated; the users access are for reading only

◆ The documents are stored in the DL and other versions can be enhanced. Moreover, the documents generally have a temporal orientation related to the publication date. So, the temporal aspect is also of interest in a DL.

◆ Finally, although a DL is not necessarily used to support management decisions, it is used to support the process of decision-making in the research. Thus, the decision-support characteristic is also of interest.

Another aspect that is important to observe is the distinction between central and local libraries, which becomes possible in the proposed approach through the differentiation between DW and DM. DW refers to the Central Library while DM refers to the Local (Departamental) Libraries.

As shown in Figure 1, in the process of the DL development (DLing), the process of a DW development (DWing) occurs as follows:

a) The data sources are not previously defined and do not consist of transactional data sources of a given company (which are generally legacy systems and relational databases). They are composed by available documents on the public or private Web instead.

b) The extraction process, instead of using conventional data extraction tools from databases or legacy archives, is made through the process of document search on the Web and its filtering. The information are accessed through traditional searching mechanisms such as

Yahoo, Altavista, Google etc., or even by the mechanisms created specially to this end which can look for documents on the hidden Web [RG2001]. An additional stage consists of the filtering of documents that are of real interest to the user.

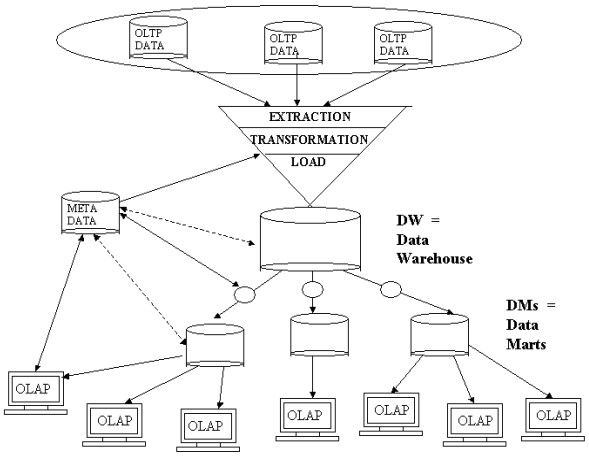
c) The transformation process consists of an analysis of the documents obtained in the previous process, capturing their metadata that are necessary to the load process of the database and will compose the library catalogue.

d) The load process, beyond effectively generating the referring catalogue of the captured documents, makes a copy of the documents found.

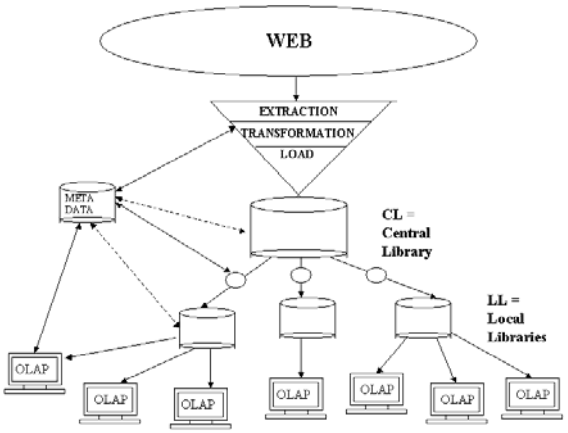
e) The DW contains documents of all the subjects of interest to a given institution, as well as their respective metadata which are organized in hierarchies according to the ontology.

f) The DM contains the documents metadata (catalogues) of all subjects relating to a department for which such database was generated.

g) The search for documents and the catalogue (DW) visualization use OLAP navigation techniques making it possible to find those whose characteristics are of interest to the user.



Data Warehousing



Digital Libraring

Figure 1 – Data Warehousing and Digital Librarying

.

3. DLing Components

Bearing in mind the parallel between DWing and Dling, the architecture components for the DL development, according to the proposed approach, are detailed below (Figure 2).

2.2.1 Data Source

The DL data source or document source is the Web, both the Public Indexable Web and the Hidden Web ([Duvillier2001] and [RG2001]). Documents can be replicated or have different versions and/or be found in various visualization formats, such as archives ”txt”, ”html”, ”ps” and ”pdf”. Considering the Web documents variety, a list of links which is part of the DL development metadata repository, assists the documents searching and allows the restriction of a set of sites for search. This list considers two types of links: links of interest (which restricts the search of the documents to these links) and undesirable links (which disregard the documents of these links in the search). Another form of restricting the universe of URLs searching is through the terms defined in the DL Ontology. In the proposed approach, the ontology is also part of the metadata repository and the DL development considers only those subjects which are of interest to the institution. The ontology initially defined can be modified manually (through the ontology owner direct manipulation) or automatically (through mappings of new subjects, that are searched by the library users and are not still part of the ontology).

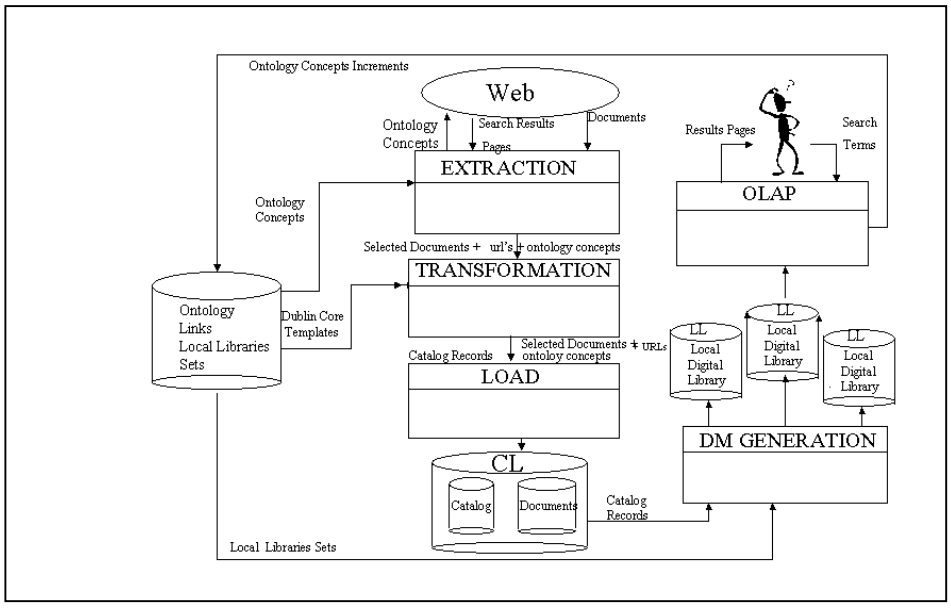


Figure 2 – Dling Architecture Components

2.2.2 – Extraction Process

A search is made initially, both on the Public and Private Web, using searching tools and traditional crawlers, such as Yahoo, Altavista, Lycos, Hobot, Excite, etc. Searching mechanisms that have been developed specially to this end and are capable of fetching documents on the hidden web can also be used. A summary of some search tools can be found in ([UC2000] and [RG2001]).

This initial search is based on the Ontology that was built for the DL and on the links list, both located in the metadata repository. Once the search result is obtained, each item is analyzed based on the abstract presented for it and according to the Ontology terms and relationships.

Based on this first analysis results, the items that seem to be of more interest are analyzed more deeply. The documents are then analyzed in more detail.

2.2.3 - Transformation Process

In this process, data needed to compose the DL catalogue are obtained. The documents that had been selected and extracted in the previous process are analyzed here. Their bibliographical information are extracted, in accordance with the standard of Dublin Core metadata [NISO2001] and the DL ontology, and then they are stored in the operational catalogue.

It is also in this process that the link mapping is updated. The document is analyzed and checked to verify whether it is already registered or not in the operational catalogue. In positive case, the document link (physical) is added to the links list. In negative case, the data are added to the operational catalogue, a logical link is created for it and the logical and physical links are added to the linklist . Each updating of the links list implies in the updating of the undesirable links list.

2.4.4 – **Load Process**

In this process, the new data of the operational catalogue, as well as their respective documents, are loaded in the DL. The load in the DL considers the multidimensional modeling for the data of the catalogue.

Documents that contain copy restriction will not be stored in the DL but their bibliographical references will be included in the DL catalogue. So, there is not a "cache" of these documents, which become inaccessible in case their original web servers are down.

2.2.5 - Digital library

In the DL all the documents whose copies had been allowed are stored, as well as their respective bibliographical references. These bibliographical references compose the DL catalogue that is modelled multidimensionally. Thus, the DL is organized based on a subject hierarchical structure and the analysis dimensions (user search) are based on the Dublin Core metadata standard.

2.2.6 - Local libraries

The local libraries store the bibliographical records of interest to a certain department. The local library catalogue is originated from a DL mapping based on a subgroup of the ontology related to the department, that will also be stored in the rmetadata repository.

2.2.7 - OLAP

The OLAP tool allows the documents retrieval from multidimensional navigations in the catalogue. A search initiates from a certain term (or set of terms). The result is presented through summaries of the documents characteristics. These summaries are quantitative graphical representations according to the analysis dimensions. The user can navigate through the dimension hierarchies, filter the result based on some characteristics and explore the Catalogue by using the navigation functions of the OLAP tools.

When the desired document is found, the user can access it through physical links (URLs). In case the site is not available or the document has been removed from the site the DL cache can be used (when there is one available).

As the user searches in a local library and does not find the documents that meets his needs, the searched terms that he used for the search are analyzed and then can be automatically included in the ontology. As such, the DL is sensible to what the users search. It is like a library that learns what can be stored in it based on the use the users made of it.

2.2.8 - Metadata

The metadata repository is composed by: the definition of the DL ontology, the links list (interesting and undesirable), the definitions of the Ontology subgroups (according to the companies/institutions) and the DL and local libraries schemas.

2.2.9 - Links Monitor

This component monitors the links of the selected objects. The links list is updated at each physical link changing.

4. DMining in Dling

4.1 - Using Data Mining Techniques in the Dling Processes

Having seen the Dwing approach for Dling, the DMining definitions as well as the use of the Dmining Techniques in Dling are presented below.

4.1.1 - Data Mining

The Internet, in particular the Web, has promoted the increase in the volume of available information. Because of this, it is increasingly more difficult to get the desired information. This calls for the use of techniques and tools which can automatically find, extract, filter and evaluate the available resources. Similarly the easiness of navigation on the Web and the competition among the various sites have generated the necessity to follow and to analyze standards of the users' accesses. In order to satisfy these needs, DMining techniques is being used.

According to [CMS1997], a natural combination of DMining and the Web, generally referred to as Web Mining, has been the focus of many papers and projects of recent research. As in any other emerging area, there is not still a formal definition of Web Mining, as it is the case of DL. Generically, Web Mining can be defined as the discovery and analysis of useful information on the Web. This information can be analysed in terms of content and use of the resources called Web Content Mining and Web Usage Mining, respectively.

Web Content Mining describes the automatic search and recovery of information and available resources located in Web sites and online databases. Web Usage Mining describes the discovery and analysis of users access standards to one or more Web services or services on line, when the Web access logs and other users navigational information are analysed.

According to [Loh1997], Web Content Mining can be divided in two sub-areas: the mining in the information retrieval (searching) and the mining in the information extraction (in the case of textual documents, it is called textual analysis) . Web Usage Mining would be the mining for discovery and analysis of access standards to the site in question.

Considering the proposed approach and the proposed architecture, we can point the use of data mining in the processes presented below**:**

Extraction

In this process, it was decided to use the DMining techniques for the ***information extraction*** in two phases: initially, when the result of the search on the Web presents a page with the links found and their respective summaries; and, later, when the text is selected, the idea is to use filtering techniques in both phases. Although the DLing approach is sufficiently generic for the multidimensional resources, in this first step of the proposed architecture development only the textual objects are being considered.

Transformation

In this process, the bibliographical records related to the selected documents will be generated using the Dublin Core metadata standard. The mining technique for the ***information retrieval*** is applicable in this stge, once the keywords to compose the metadata tags will be extracted from the documents. The main information extraction techniques recognize the structures of a text (data or information) through the tag analysis, generally represented by keywords.

OLAP

In this process, it is expected that the results of the searchs in the library are analyzed using mining techniques for the ***information discovery***. When a term analysis is not in the ontology and it indicates a new way of research, that is, an interesting aspect to be included in the ontology, this new term and its respective relationships with the Ontology existing terms will be included in the ontology.

5. Conclusion and Future Work

In this paper we proposed a new approach to the development of digital libraries. Since the main issues in the DL development are complex data integration and difficult information retrieval, we proposed the development of the DL based on the Data Warehousing approach, since it is being efficiently used in the integration of complex data and in the assistance in the information retrieval. We correlated the Dwing and Dling processes, described each of their componentes and defined the DW as being the Central Libraries while de DMs as being the Local Libraries.

Another research area that has been solving information extraction, retrieval and discovery issues is the Data Mining area. As such, the data mining techniques are useful and applied, in three of the Digital Librarying processes: Exctraction, Transformation and OLAP. These techniques will allow the filtering of the really interesting documents in the Extraction Process, the generation of the Dublin Core Metadata, using the keywords found in the selected objects, and, finally, and automated refresh of the library, discovering the users needs through the analysis of their accesses standards.

As we mentioned in the beginning of this paper, in this work a DL is considered a large object collection, in diverse digital formats, persistent, managed and well organized using a catalogue and accessed through the Web. Initially we restricted the scope of the DL to textual objects only. Now, after reviewing our proposal and being involved with the PGL Project, we are suggestion the use of this approach to LOs. It fits very well to sort out the access problem to the various LO-DL (Learning Objects Digital Libraries) which will compose our project as well as those found on the Web.

Our goal is be to integrate the various LOs repositories or PGL LOs Digital Libraries and/or others, through the diverse metadata standards integration, thus building an unified DW. .

The DW, or DL, would become LO-DL (Learning Object Digital Library), containing the Catalogue (in Dublin Core Metadata) of all the LOS that are of interest to our Project. DMs would be created in accordance with the needs of the users´ searches. An example of a DM (local LO-DL) could be a Catalogue pointing to LOs written in Portuguese and for children of a certain age range..

Bibliography

ACDFF+1995 Andresen, D., Carver, L., Dolin, R., Fischer, C., Frew, J., Goodchild, M., Ibarra, O., Kothuri, R., Larsgaard, M., Manjunath, B. S., Nebert, D., Simpson, J., Smith, T. R., Yang, T., Zheng, Q.; "The WWW prototype of the Alexandria Digital Library", In the Proceedings of

ISDL'95: International Symposium on Digital Libraries, Japan, 1995, http://www.dl.ulis.ac.jp/ISDL95/proceedings/pages75/17.html

Chen2000 Chen, H.; "The Illinois Digital Library Initiative Project: Federating Repositories and Semantic Research", 2000, http://ai.bpa.arizona.edu/hchen/docs/DLI/

CMS1997 Cooley, R., Mobasher, B., Srivastava, J.; *Web Mining: Information and Pattern Discovery on the World Wide Web*, 1997, http://www-users.cs.umn.edu/~mobasher/webminer/survey.html

Duvillier2001 Duvillier, L.; *The Hidden Web*, 2001, http://europa.eu.int/comm/development/ publicat/courier/courier_184/ en/en_070_ni.pdf

GAAS1999 Geffner, S., Agrawal, D., El Abbadi, A., Smith, T.; *Browsing Large Digital Library Collections Using Classification Hierarchies*, In Proceeding of the CIKM'99, 1999

Inmon1996 Inmon, W.H.; *Building the Data Warehouse*, John Wiley & Sons, Inc., 1996

Kimbal1996 Kimball, R.; *The Data Warehouse Toolkit*, John Wiley & Sons, Inc., 1996

Loh1997 Loh, S.; *Knowledge Discovery in Text Databases*, 1997, http://www.ulbra.tche.br/~loh/apostilas/dc-texto.html (in portuguese)

NISO2001 NISO; *The Dublin Core Metadata Element Set,* An American National Standard, National Information Standards Organization (approved by the ANSI in September, 2001)

PBCCG2000 Paepcke, A., Baldonado, M., Chang, C.K., Cousins, S., Garcia-Molina, H.; "Building the InfoBus: A Review of Technical Choices in the Stanford Digital Library Project", 2000, http://dbpubs.stanford.edu:8090/pub/2000-50

RG2001 Raghavan, S., Garcia-Molina, H.; "Crawling the hidden Web", . In the Proceedings of the 27th Intl. Conf. on Very Large Databases (VLDB), Italy, 2001

RG2001 Raghavan, S., Garcia-Molina, H.; *Crawling the Hidden Web*, 2001, http://dbpubs.stanford.edu:8090/pub/2001-19

Schwartz2001 Schwartz, C.; "LIS 462 – Digital Libraries Definitions", 2001, http://web.simmons.edu/~schwartz/462-defs.html

UC2000 UC Regents, *Internet Search Tool Details*, 2000, http://sunsite.berkeley.edu/Help/searchdetails.html

WB1998 Weinstein, P., Birmingham., W.; “Organizing Digital Library Content and Services with Ontologies”, International Journal on Digital Libraries, special issue on artificial intelligence for digital libraries, 1998