Proceedings of the 10th WSEAS International Conference on COMPUTERS, Vouliagmeni, Athens, Greece, July 13-15, 2006 (pp224-229)

Using Data Mining for the Refresh of Learning Objects Digital Libraries

CÁSSIA BLONDET BARUQUE, RUBENS NASCIMENTO MELO
Computer Science Department
PUC-Rio
R. Marquês S. Vicente, 225, RJ
BRAZIL

Abstract: - The LO-DL (Learning Objects Digital Library) Project is being developed at PUC-Rio in the Database Tecnology Lab (TECBD). This Project aims at integrating LOs repositories through their metadata in a uniform catalog or Digital Libray (DL), making it transparent to the users their locations and characteristics. The process of digital library development includes issues such as the integration of several databases. Moreover, access to the DL (Digital Library) must be assisted by the use of content hierarchies that guide the user in the discovery and filtering of information of his/her interest. In this work we propose a new approach for developing Digital Libraries of Learning Objects using a Data Warehousing Architecture, which is a method that addresses both issues mentioned above. We make comparisons between the components and services of both the Data Warehousing and the "Digital Librarying" Architectures. Furthermore, we suggest the use of Data Mining Techniques in some steps of the building and utilization of the DL. In particular we will detail the users profiles´ analysis process which uses the library access log and a data mining tool for the library automatic refresh. We propose the use of association rules for the detection of the users needs so that the system can then search and make the new LOs available in the next loading (refresh) of the library. The supporting database which contains the Ontology (in our case a Taxonomy) of the LOs of interest is also described in the paper.

Key-Words: – digital library, data warehousing, data mining, learning objects, e-learning..

1 Introduction

Learning Objects (LOs) are being developed using diverse tools (Flash, Dreamweaver, Frontpage) as well as various languages (HTML, Java, etc), and, recorded on the Web, based on metadata standards. A number of different metadata standards are used by the course developers so that the final users must query different Web sites. Furthermore, each LO repository requires from the user a different way to access it in such a way that he needs to get used to various interfaces.

In this context, our proposal to develop **educational digital libraries** (or LOs libraries) aims at integrating LOs from multiple sources through the use of a well established database metodology, i.e, Data Warehousing (DWing), so that they can be accessed through a single interface, while making it transparent to the user their location (repository) as well as the metadata which describe them (Dublin Core, LOM, METS, etc).

Since LOs are distributed all over the Web as well as their catalogue, it is difficult for the user to search and find the LO he needs. Furthermore, once it is located, as it happens to other documents on the Web, it is difficult to define precisely whether its content meets the user´s requirements based only on its content description, i.e. its metadata.

As such, there is a need for the homogenization of object descriptions. This calls for the use of a common metadata standard together with a common metadata´s subject classification and also the integration of these descriptions (catalogues), so that it looks like there is only one single catalogue to the user or searcher.

One database research area which has definitely been contributing to solve issues related to complex databases and data integration is DWing (DWing).

The digital libraries (DL) development process involves issues such as integration of complex data found on the Web. Additionally, the access to digital libraries should be supported by the use of content hierarchies which guide the user in the finding and filtering of information relevant to him/her.

This work presents an architecture for the development of DL based on the DWing approach (methodology) and using Data Mining techniques (DMing). DWing has been used for data integration in the context of decision-making processes. As a result of DWing, there is a database called Data Warehouse (DW).

This is subject-oriented, it promotes content organization in hierarchies and in an integrated fashion. DMining, on the other hand, allows the location, extraction, filtering and evaluation of the desired information and digital objects, as well as the tracking and analysis of users´acess patterns.

Since DWing has well been addressing issues such as complex data integration and easiness in access to data, we proposed in [1,2] its use to promote the integration of LOs located in various repositories on the Web and to make access to them easier.

The resources organized in a multidimensional fashion, following the DWing modeling,, can be queried in a number of ways, from different perspectives, enhancing the diversity of traditional query types found generally in libraries, which include subject, title, author and key-words. Bearing in mind the various dimensions that would be the different "tags"´contents (metadata elements), the queries can be combined and, therefore, one could query authors by periods of time, for certain specific subjects and so forth.

OLAP technology can thus greatly contribute to enhance digital library queries and, consequently, to improve Web-based education (WBE). Digital libraries are, in reality, an integral part of WBE, not only from the perspective of the educational resources´ producers but also from the consumers´s perspective. As such, teachers, instructional designers and learners can all benefit from it, since they need to locate and build content on-line in an easy way.

Recent researches have focused on the construction of Web-based educational digital library [3]. As for educational applications, one of the main objectives of DL is to provide facilities for the search and use of LOs, so that cataloguing systems are developed to give the semantics of LOs The metadata structures are easily found and, thus, they allow the finding, sharing and reuse of LOs. Some educational digital libraries emerging on the Web – **SMETE** [4]**, ILUMINA** [5] e **DILLEO** [6] **-** emphasize the LO orientation and are, therefore, similar to our work.

The rest of this paper is organized as follows: In **Section 2**, we give an overview of our proposal for the development of digital libraries; in **Section 3,** we describe how to apply data mining techniques on the Library users' access LOG in order to improve its quality providing the users with what they need. Finally, in **Section 4** we present our concluding remarks.

2 The LO-DL Development Architecture

Learning Objects (LOs) are being developed using diverse tools (Flash, Dreamweaver, Frontpage) as well as various languages (HTML, Java, etc), and, recorded on the Web, based on metadata standards. A number of different metadata standards are used by the course developers so that the final users must query different Web sites. Furthermore, each LO repository requires from the user a different way to access it in such a way that he needs to get used to various interfaces.

In this context, our proposal to develop **educational digital libraries** (or LOs libraries) aims at integrating LOs from multiple sources through the use of a well established database metodology, i.e, Data Warehousing (DWing), so that they can be accessed through a single interface, while making it transparent to the user their location (repository) as well as the metadata which describe them (Dublin Core, LOM, METS, etc).

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.

In accordance with William H. Inmon [7], a DW is a collection of data that is subject-oriented, integrated, non-volatile, variant in the time and used for supporting management decisions.

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 LOs through a subject hierarchical classification).
- a DL must have an integrated view of LOs. A possible distribution of these LOs, 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 LOs are stored in the DL and other versions can be enhanced. Moreover, the LOs 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.

2.2 DLing Components

Figure 1 shows the parallel and the differences between the DWing (Data Warehousing) and the DLing (Digital Librarying) architectures. Bearing in mind this parallel between DWing and DLing, the architecture components for the DL development, according to the proposed approach, can be read in [2]. They are:: data source, Extraction Process, Transformation Process, Load Process, Digital Library (DW), Local Digital Libraries (DMs), OLAP, Metadata and Links Monitoring.

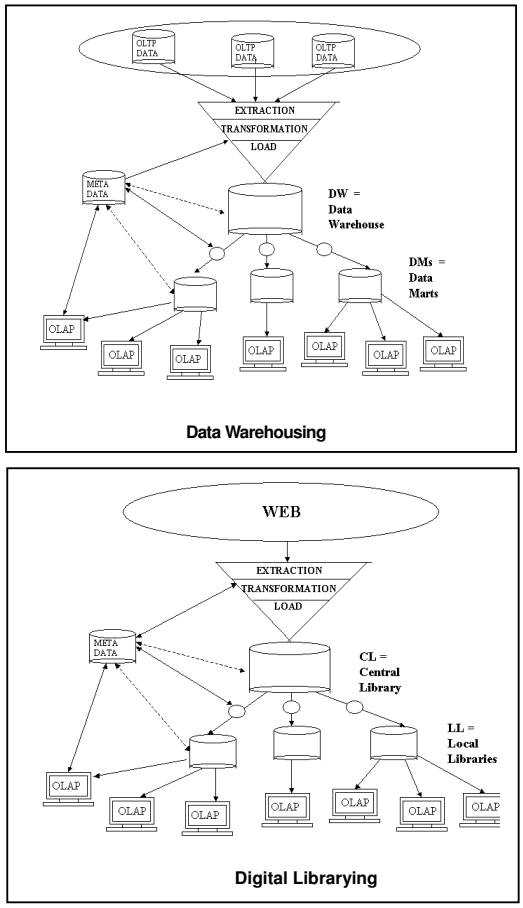

FIGURE 1 – DATA WAREHOUSING AND DIGITAL LIBRARYING

2.3 IDLing: Using Data Mining Techniques in the DLing Process

In our proposal, data mining techniques [8] are incorporated in the DWing architecture to assist in the automation of some processes and in making more "intelligent" and sophisticated analysis available on the library contents and their use. It is the process of the improvement of the Library content that we will describe in the next topic. The other processes that use data mining techniques are not described in this paper´and can be found in [1].

3 Data Mining and the Automated Refresh of the LO-DL Library

In this process the users´ accesses to the Library will be analyzed in order to detect whether there is a change in their profile. Since our goal is to improve the Library quality, which will allow to better meet the users´ needs, so when this change is detected an automatic update ("refresh") of the Library will be run. For this reason, it is necessary to analyze this profile change, which will be made through the use of data mining techniques.

3.1 Refreshing the Library

The search to the library asset is OLAP-like [1]. OLAP allows a variety of combined searches to the Library. One of the possibilities when searching for LOs is using the "subject" dimension. In order to facilitate the understanding of this process, we will consider "subject" as being an index without topic hierarchies and related terms, in a different way of that presented in the proposed ***Ontology*** [1].

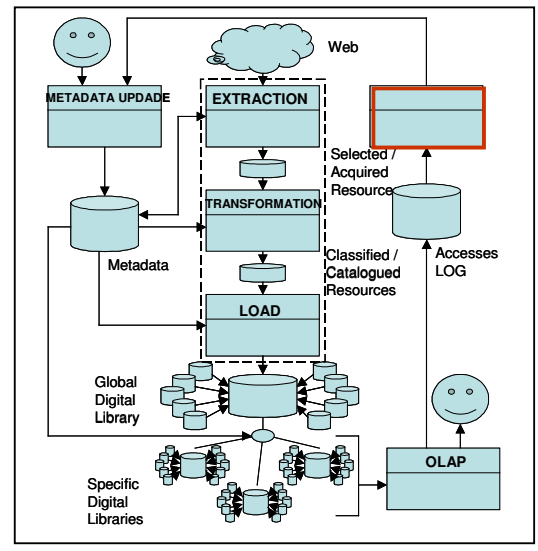

FIGURE 2 – IDLING AND THE PROCESS OF QUERYING AND ANALYSING " À LA OLAP "

It is necessary to register such searches or queries in a Library Access LOG. Information like the user identification (for privacy care substituted by a codification), the subject being used as the search key to the Library and its result, informing whether LOs pertaining to the subject searched were found or not will be recorded in this LOG. This process will be executed periodically (weekly) since the analysis is made considering these accesses occurrences amount.

The possibility of users profile changing will exist when they start searching for LOs of a subject that the Library does not contain. Such searches, already recorded in the LOs accesses LOG, will be analyzed and, if it is verified a user profile change, the new subjects which will be considered interesting, based on this analysis, will be included in the *Ontology* database, which will trigger the automatic "refresh" of the Library.

Certainly searching LOs by subject that do not exist in the Library is not suficient to determine a users profile changing. The changing will only happen when there are associations between previous and new searches made by these users. In orde to analyze these associations, data mining techniques that allow discovering association rules among data will be used. Only after this discovering the decision will be made – will the subject be included or not in the Library.

Association rules is the name of one of the different techniques used in the KDD (Knowledg Discovery in Databases) process in order to obtain knowledge. KDD is the process, which begins with data preparation, through the cleaning, categorization, transformation, etc, until these data can be submited to the data mining process which provide results that should be still analyzed by human. Since our data, in this process, have already suffered all the previous steps, they are ready now for the data mining process.

In [9, 10] this technique is described in detail, with the specification of related terms and its application in real examples. We summarize below the main aspects applied to our problem.

Association Rules represent items combination or attribute values that occur on a certain frequency in a database. A typical application is the pattern analysis of the purchases made by consumers. The technique applied in the consumers database generates, as a result, a set of rules that determines which products are acquired together. Based on this result, one can decide to put these products close to the others on the market shelves.

The **Rules Association** technique, differently of what happens in traditional statistical methods, does not require that the specialist tests its hypothesis against the database since it is a technique directed to the information discovery, which extracts automatically hypothesis and patterns from their databases. In our work the goal is to discover rules that show that some new subjects, which are being searched, deserve to be included in the Digital Library while others do not.

An Association Rule is an expression as A ➔ B, where A is the antecedent and B is the consequent of the rule, what means that the probability of B occurrence increases when A also occurs. Both the antecedent and the consequent of a rule can contain one or more items. The Association Rules can be transactional, multidimensional, hibrid or multilevel). They are generated to be analyzed based on the measures such as support, confidence and interest.

Transacional Association Rules are those that contain only one atribute (or dimension) in the antecedent and only one attribute in the conseqüente, and, in both sides the attribute is the same. The **Multidimensional** ones allow having a set of attributes in the antecedent and a set of attribute in the conseqüent, but the attributes cannot be repeated in both sides. The **Hibrid** were developed to solve these issues, and so, they are like the Multidimensional but allow the attribute in the antecedent to be repeated in the consequent.

Support, confidence and interest are the basic measures used in the analysis [9, 10]. Given an A ➔ B rule its **support** measure represents the percentage of the database transactions that contain the items A and B, indicating the relevance of the rule. The **confidence** measure represents, among the transactions that contain A items, the percentage of the transaction that contain also the B items, indicating the validity of the rule. The **interest** measure represents the how much frequently turns to B when A occurs. This last one is computed in following way: **Interest (A ➔ B) = Conf(A ➔ B) / Sup (B).**

There are many data mining tools that use the association rules technique. However the scope of the attributes, which they allow in each side, is variable. There are tools that allow the choice of only one attribute for the conseqüent, which is the case of the tool we are using in our application and that is considered adequate in our case. There are no tools that process the hybrid association rules case?, and then, it is necessary to adopt some tricks in order to run the tool with this goal.

MagnumOpus is the name of the tool we are using in our wor*k*. It is a commercial tool but we are using the trial one which encompasses 1.000 cases analysis. In [11, 12] there are detailed descriptions on how to use this tool. Some techinal publications, in which the

technologies used in the tool are described, can be read in [13, 14, 15, 16 and 30].

Before using the tool it is necessary to import the database on which the analysis will be done. The Magnum Opus tool uses two entry files: the first contains the database schema description informing whether the attributes are categoric or numeric. If the attribute is numeric it is also necessary to inform in how many categories the tool has to transform it. The data file must be imported after the schema. This file must be stored in a "txt" file format.

Prior to importing the database it is still necessary to transform the accesses LOG generated during the OLAP process – during the Library query.. The field "subject" and the field "that indicates whether the search was successful or not" should become a single field. The subjects which were not found in the Library have to be prefixed with "NF" and the ones that were found with "FO". Then the tool will import a file that contains only two fields: the user id and the subject.

By applying the tool in our case, we should import the schema, which includes the user code (instead of the user id in order to keep its privacy) and the subject, and then, import the LOG data which had already been changed. Next step is to decide what type of association rules to generate: the transactional or multidimensional.

In the case of our work, we need to discover association rules between subjects. As such, we have to choose the attribute (or dimension) SUBJECT as both antecedente? and descendente?, what takes us to the *transactional* association rules, but we have an issue to solve: besides the SUBJECT we need to use other field to make this analysis, which is "Does the SUBJECT exist in the Library or not?". This would take us to the choice of one more attribute and we would have to use hybrid association rules.

As we mentioned before, there are no hybrid association rules tool. As such it is necessary to use a trick that allows us to use the transactional association rules tool as being hybrid. So, before importing data to the tool it is necessary to transform them. For the SUBJECTs, which were searched and were found in the Library we will include an "FO" prior to the SUBJECT name. For the cases where the SUBJECT was not found, we will include a "NF" prior to the SUBEJCT name.

So, applying the tool the our case the right choice is the generation of the association rules of the transactional type, since what we need is to associate subjects between themselves, and this type of rule referes to a same attribute being repeated in both sides, the antecedente and the consequente. The rules options to be generated are: "FO-subject" ➔ "NF-subject", "FO-subject" ➔ "FO-subject", "NF-subject" ➔ "FO-subject" and "NF-subject" ➔ "NF-subject".

The third step is the choice of the attributes which should participate on the association rules generation process, by selecting, among the attribute of the database, those which will compose the antecedents side and those which will compose the consequence side of the rule. The tool shows the possible attributes in two windows. The left one is related to the antecedent and the right one to the consequent.

The next step is to choose the measures which are the basis for the searching of rules. The tool allows the searching of rules using interest, confidence, support, and others measures. Values are assigned to these measures. The tool will exhibit the rules whose values are greater than the ones assigned to interest, confidence and support. To choose the better values to be assigned to these measures, it is necessary to test the tool with the database. .

Having had made all the choices, the tool can now be started. The tool generates an output containing a list of all the rules that were discovered and have their values greater than the values set to the measures. Having discovered which rules determine their data association, the administrator can, then, make a decision in order to improve its business. In our case this decision is automated – the new subjects will be included in the ***Ontology*** table.

Finllay, the rules discoverd by the tool which contain both in the antecedente and the conseqüente a subject with the prefix "NF" should be considered to the process of the inclusion of the respective subjet in the Ontology. As mentioned before, the Ontology contains the subjects that compose or will compose (the search for LOs in the Internet is based on it). Once these new subjects are included in the Ontology, the process ***Index and Update the Digital Library*** is executed autocamically, resulting in a new version of the Library, which will contain LOs pertaining to the new subjects included in the Ontology.

It is important to note that in our work we use the Ontology to analyse the users profile changing, and, consequently, a refresh of the library is made automatically. This idea can be applied to any other dimension (attribute) of the Library and as an example we could be trying to find out rules regarding the format of the LO. Users could be used to search for textual LOs in the Web. With the technology advances they could start to search for video LOs, what changes its profile.

4 Conclusions and Future Work

In this paper we made an overview of the DL development approach proposed in [1]. The LO-DL Project is being developed at PUC-Rio in the Database Technology Lab (TecBD). We detailed the "OLAP" component of the proposed Digital Librarying Architecture and presented some users' queries and managerial analysis scenarios.

Once 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 a well stablished technique to address these issues. We made a comparison between the Dwing and Dling processes, described each of their components, and, defined the DW, as being the Central Libraries, while the DMs, as being the Local Libraries.

We concluded that viewing a DL as a DW is very appropriate not only because the characteristics of a DW can be well applied to a DL, but also because it allows a much more sophisticated way of searching its content, different from the usual searches by author, subject and title. The navigation through the hierarchies of contents is also an important aspect facilitated by our approach. Another advantage that comes from this approach is the process of decision-making. OLAP technique helps users not only in finding the content they need, but also in the managerial process of the DL. The DL can be analyzed so that the managers or librarians can decide, for example, in what kind of LOs they should invest more or less money. The DL analysis can detect users' needs and meet them in an automatic and efficient fashion, thus improving the library quality. This process which may be supported by data mining techniques was described in detail in this paper.

References

[1] Baruque, C.B: Desenvolvimento de Bibliotecas Digitais de Learning Objects Utilizando Técnicas de Data Warehousing e Data Mining Tese de Doutorado, PUC-Rio, 191pp., 25,Abril,2005

[2] Baruque, C.B. and MELO, R.N. Using Data Warehousing and Data Minining Techniques for the Development of Digital Libraries for LMS. Proceedings of the IADIS International Conference. Madrid, Spain. 6-9 October 2004.

[3] Vladoiu, M.. Learning Objects need badly Instructional Digital Libraries Support; Informatics in Education International Journal Vol. 2, No. 2, pages 291-316.,2004

[4] Muramatsu, Brandon; Manduca, Cathryn A; Mardis, Marcia; Lightbourne, James H; McMartin, Flora P. Panel: The National SMETE Digital Library Program, Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries**,** Virginia, United States, p.p-278 – 281,2001

[5] ILUMINA. http://www.ilumina-dlib.org/. Accessed in April 2006.

[6] Mikulecka, J, e-DILEMA – Socrates Project No 90683-CP-1-2001-Minerva, University of Hradec Králóve, Czech Republik. Roma, september 2003. http://e-dilema.uhk.cz/doc/FinalWeb.ppt Accessed in : April 2006.

[7] Inmon, W.. Building the Data Warehouse. John Wiley & Sons, Inc..1996.

[8] 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

[9] Gonçalves**,** E.; Plastino, A..**Mining Strong Associations and Exceptions in the STULONG Data Set**. Proceedings of the ECML/PKDD 2004 Discovery Challenge. Pisa, Italy, p. 44-5. September 2004.

[10] Gonçalves**,** E..**Mineração de Exceções Negativas e Positivas**. Dissertação de Mestrado. Universidade Federal Fluminense, 2004.

[11] G I WEBB & ASSOCIATES**. Using Magnum Opus with transaction data: a tutorial introduction.** http://www.rulequest.com/MOUsingBasket.html. Acesso em: março de 2005.

[12] G I WEBB & ASSOCIATES**. Using Magnum Opus with attribute-value data: a tutorial introduction**. http://www.rulequest.com/MagnumOpus-win.html. Acesso em: março de 2005.

[13] Webb, G. **Efficient Search for Association Rules**. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 99-107. ACM Press, 2000.

[14] Webb, G.; Zhang S. . **Beyond Association Rules: Generalized Rule Discover**y, Submitted for publication. http://www.csse.monash.edu.au/~webb/Files/WebbZhang03.pdf. Acesso em: março de 2005. 2003.

[15] Webb, G.; Butler A.; Newlands D.. **On Detecting Differences Between Groups**. Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03), pages 256-265. ACM, August, 2003.

[16] Webb, G..**Preliminary Investigations into Statistically Valid Exploratory Rule Discovery**. Proceedings of the Australasian Data Mining Workshop (AusDM03), pages 1-9. University of Technology, Sydney, 2003.

[17] Webb, G.. **OPUS: An Efficient Admissible Algorithm For Unordered Search**. Journal of Artificial Intelligence Research, 3:431-465. 1999.