Data Mining as a Tool for Information Retrieval in Digital Institutional Repositories

Leticia Tonon
Department of Information Systems
University Center Euripides Mar Ília / UNIVEM
Mar Ília, Brazil
e-mail: le.tonon@univem.edu.br

Elvis Fusco
Department of Information Systems
University Center Euripides Mar Ília / UNIVEM
Mar Ília, Brazil
e-mail: fusco@univem.edu.br

***Abstract*— Currently there is a large volume of data stored in informational bases of digital repositories and the problem of finding useful data in information retrieval systems has intensified, making the processes of recovery increasingly sophisticated. This study aims to make use of data mining techniques to improve information retrieval in digital institutional repositories.**

Keywords- institutional digital repositories; information retrieval; data mining

I. INTRODUCTION

With the intense use of computational means for creating scientific productions and general purpose documents a large amount of information has been generated but leaves a lot to be disseminated to those people and institutions interested on it and also end up being lost and need to be redone, which could easily be reused thus providing time and personnel savings in organizations (9).

Such information needs to be stored so they are not lost in time, so they can be accessed by interested parties and that are preserved in order to enable the reuse of documents created by the institutions. Thus, it is necessary to use systems capable of fulfil these needs which are present in digital repositories, systems able to store, manage, preserve and disseminate the productions of any institution (6).

The great increase of storage capacity of computers and the evolution of the Internet and networks have made possible to think in digital repositories capable of allowing access to the full content of works in digital format. Barton (2) defines institutional repositories as "a database with a set of services to capture, store, index, preserve, and deliver research of an educational institution in digital formats."

This paper aims to extend the process of Information Retrieval in Digital Environments Informational Institutional Digital Repositories by Data Mining techniques aimed at increasing the relevance and effectiveness of search results information.

II. INSTITUTIONAL DIGITAL REPOSITORIES

Digital repositories are collections of articles and files on several different extension, but digital, available to be accessed through computational means, can be accessed from local networks or by Internet. There are no limitation for contents on digital repositories, they may be institutional such as research projects group, graduate theses, several studies developed by members of the institution, monographs and also repositories for any type of files in digital format desired (8).

In the case of institutional repositories of universities, Lynch (7) defines as a set of services that an educational institution offers members of the community aiming at efficient management and dissemination of digital materials created by the institution and by members of their community.

III. DSPACE

The Libraries are working to extend their services into the digital era, to reflect current trends in scholarly communication and education, and to offer new means of distributing research material that are enabled by network technology. It is incumbent upon libraries to develop strategic and economic plans for the preservation and usability of those resources over time. Analogous to print materials, digital library initiatives require cost-benefit considerations that must be carefully weighed against other library priorities (3).

The DSpace is a free software widely used for the development of institutional repositories, was developed by the Massachusetts Institute of Technology (MIT) and Hewlett-Packard Laboratories.

Its structure provides a model of organizational based in "communities" and collections, which can be configured to reflect the full range of administrative units of an institution. Allows configuration of the editorial process in the mold of traditional journals, including the possibility of peer review. Supports all kinds of digital file formats, including text, sound and image (11).

IV. INFORMATION RETRIEVAL

Information retrieval is the area that deals with the automatic storage and retrieval of documents.

An Information Retrieval System has three basic components: Acquisition and representation of the information need; Identification and representation of the document content, and; Specification of the comparison function that selects relevant documents based on the representations(8).

The most important tool to aid the recovery process is called an index, which is a collection of terms that indicate the place where the desired information can be located (5).

Scholtes (10) claims that information retrieval requires the joint application of techniques for Natural Language Processing and Artificial Intelligence. In general all information system is a system that retrieves information. One of the first types of institutions that have adopted information systems were libraries, at first these systems only either automated search process. With time came to use probabilistic models so that queries to submit relevant content.

The Boolean, vector and probabilistic models, classically used in information retrieval process are used to form the search queries that are relevant to the query content.

Most information retrieval systems of digital repositories allows the user to express what information it needs and from that the system returns the documents deemed relevant. Studies in information retrieval may involve several aspects involving the conceptual universe and sociolinguistic diversity.

V. DATA MINING

In the information retrieval process, highlight the process of Data Mining, which according to Diniz Neto (4) is "the process of extracting information without prior knowledge of a large database, and its use for decision making."

Before of knowing the process of data mining is necessary to know the definition of data, information and knowledge in level of interrelation, illustrated in Figure 1.

Figure 1. Architecture Proposal Information Retrieval

Data are elements in their raw form, in the other words, most are not understandable; Information is the result of process of treatment of data, namely, those not understandable data are transformed into something that can lead to understanding; Knowledge is the result of the analysis of information, that is, when the information starts to be understood and used to optimize any existing process in the system.

In a simplified way, the process of data mining starts from cleaning a data source where, for example, are taken from noise and redundancy. Then, these data go to Data Marts and/or Data Warehouses. Finally, using the repositories you can select columns to enter the data mining process. At the end of the process the information is finally generated (in form of charts, spreadsheets, etc.) analyzed and become knowledge to the user.

It should be noted that each data mining technique or each specific implementation of algorithms used to conduct operations data mining adapts to certain problems more than others, which shows that there is no universal method of data mining better. To each particular problem has a particular algorithm. Therefore, the success of a data mining task is directly linked to the experience and intuition of the analyst (4).

The goal of content personalization is to ensure that the right person receives the right information at the right time (1). With the methods of Data Mining would be possible to generate specific profiles for each user group, to thereby make the personalization of information retrieval processes possible.

VI. DEVELOPMENT

So far the literature sources and survey related work was performed, as well as a study of the techniques of Data Mining which enabled check for various tasks Data Mining where for each one there are specific techniques to be used. To the accomplishment of the Data Mining in the Weka API design, widely used for research purposes, tools of their algorithms can be applied directly or tools used in Java applications will be used.

To the research, an institutional repository that uses DSpace as tool and Dublin Core as metadata schema will be used. To collect the metadata OAI-PMH, which is a mechanism for collecting metadata repositories will be used. This protocol consists in six verbs each one of them behind a specific type of XML response. ListRecords was defined as verb this verb extracts the metadata from the repository and with the help of optional arguments you can make a selective extraction based on Dublin Core metadata elements.

The development of a Java application started in NetBeans IDE, this application will be responsible for liaison between Information Retrieval and Data Mining.

VII. PRELIMINARY RESULTS

Based on theoretical considerations and development ever conducted was possible to verify that the concepts of data mining can be used in the information retrieval process in digital repositories, according to the proposal presented in Figure 2.

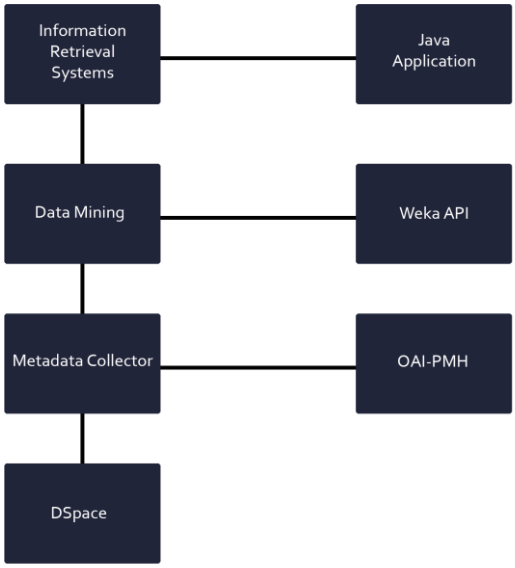

Figure 2. Architecture Proposal Information Retrieval

The architecture proposes that a layer of Information Retrieval is implemented using techniques of Data Mining in Digital Repositories. This layer is in the collecting metadata, so you can later go through the Data Mining techniques with the aid of the Weka API Java application.

Given the proposed architecture, the layer of extracting metadata from DSpace server has already been developed. For this it is necessary to inform: the server on which you want to extract the metadata, the OIA-PMH command, which in this case is the ListRecords, the period you want to view (if the fields "From" and "To" are not filled a link will be generated with all the metadata server).

The partial result of this query can be seen in Figure 3. Due to the length of answer some records were omitted.


```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2014-01-30T17:58:00Z</responseDate>
  <request from="2011-12-01T00:00:01Z" metadataPrefix="oai_dc" until="2013-11-12T08:00:01Z"
  verb="ListRecords">http://aberto.univem.edu.br/oai/request</request>
 ▼<ListRecords>
  ▶<record>...</record>
  ▶<record>...</record>
  ▶<record>...</record>
  ▶<record>...</record>
  ▼<record>
   ▼<header>
     <identifier>oai:aberto.univem.edu.br:11077/244</identifier>
     <datestamp>2012-10-15T13:56:16Z</datestamp>
     <setSpec>hdl_11077_230</setSpec>
    </header>
   ▼<metadata>
    ▼<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
     http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
     ▼<dc:title>
        Jornal da Fundação - Edição nº 149 - Ano XVI Maio/2011
      </dc:title>
      <dc:creator>UNIVEM</dc:creator>
      <dc:date>2012-10-11T13:32:00Z</dc:date>
      <dc:date>2012-10-11T13:32:00Z</dc:date>
      <dc:date>2011-05-01</dc:date>
      <dc:type>Periódico</dc:type>
      <dc:identifier>http://hdl.handle.net/11077/244</dc:identifier>
    </oai_dc:dc>
   </metadata>
  </record>
 ▼<record>
   ▼<header>
     <identifier>oai:aberto.univem.edu.br:11077/245</identifier>
     <datestamp>2012-10-15T13:56:14Z</datestamp>
     <setSpec>hdl 11077 230</setSpec>
```


Figure 3. Response Metadata Extractor

VIII. FINAL THOUGHTS

Nowadays most of information retrieval systems in digital repositories are not using Data Mining techniques for personalized search. The use of this type of search retrieval systems provide more dynamic information in digital repositories that would be capable of bringing more relevant results to the user by means of discovered patterns in databases.

In this context, the problem of identifying and retrieving which data mining techniques can be used in this scenario and list the advantages and disadvantages of using them.

As important as the storage of information in Digital Repositories, are the processes of Information Retrieval, which cause users to have access to the data compiled for decision making and to generate new knowledge.

In this sense, data mining and its artificial intelligence algorithms collaborate in the process of information retrieval in digital institutional repositories.

REFERENCES

[1] ARANHA, Francisco. Análise de Redes em Procedimentos de Cooperação Indireta: Utilização no Sistema de Recomendações da Biblioteca Karl A. Boedecker. São Paulo: EAESP/FGV/NPP, 2000.

[2] BARTON, M. R. Creating an institutional repository: LEADIRS workbook. Cambridge-MIT Institute, 2005. Disponível em: <www.ugr.es/~afporcel/construccion.pdf>. Acesso em junho de 2013.

[3] BARTON, M. R.; Walker, J. H. Building a Business Plan for DSpace, MIT Libraries' Digital Institutional Repository, 2003.

[4] DINIZ, Carlos Alberto R., NETO, Francisco Louzada. Data Mining: uma introdução. São Paulo: Associação Brasileira de Estatística, 2000. 123p.

[5] FRAKES, W. B. & Baeza-Yates, R. Information Retrieval Data Structures & Algorithms, Prentice Hall, 1992.

[6] LEWIS, S.; YATES, C. The DSpace Course - Introduction to Dspace. CADAIR, 2008. Disponível em: <http://cadair.aber.ac.uk/dspace/handle/2160/617>. Acesso em junho de 2013.

[7] Lynch, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. Association of Research Libraries, n. 226, 2003. Disponível em: <www.arl.org/bm~doc/br226ir.pdf>. Acesso em junho de 2013.

[8] ROMANI, Lucas Salviano. Análise e Implantação de Repositório Digital utilizando Software Livre DSpace. 2009. 98f. Trabalho de Conclusão de Curso de Bacharelado em Ciência da Computação. Centro Universitário Eurípides de Marília, 2009.

[9] SCHOLTES, J. C. Neural Networks in Natural Language Processing and Information Retrieval. PhD thesis, Institute for Logic, Language and Computation (ILLC).University of Amsterdam, 1993.

[10] VIANA, C. L. M. Repositórios institucionais baseados em DSpace e ePrints e sua viabilidade nas instituições acadêmico-científicas, 2006.