International Accesses to a Digital Library of ETDs

Ana M B Pavani
DEE, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil

Keywords: Access, logs, statistics.

ABSTRACT

This paper addresses the analysis of accesses to a digital library of ETDs using the information on server logs and metadata on the database. The first part describes the method used to gather and store data – filtering the logs of the Apache Server, grouping info and modeling the database to store it. The second presents the basic functions used to analyze data – focus on authors, research areas, graduate programs, countries and time frames. Combinations of basic functions are described, as well as the characteristics of the model – functions are processed online and in real-time; current month data is time-varying; the model treats data as sequences of points with discrete-time intervals of a month.

1. INTRODUCTION

The creation an ETD digital library aims at making ETDs more accessible and utilized. At a first thought, faculty and graduate students of the university are the potential users. At a second glance, the fact that t digital library is connected to the Internet indicates that horizons may be further away. Then, there are union catalogs and metadata harvesting; the world seems to be the limit.

The objective of this work is to present a preliminary analysis of the accesses to the ETDs on the digital library of the Maxwell System ("Maxwell"). The digital library is an OAI-PMH data provider, and harvested metadata are made available from union catalogs and other portals. Some ETDs are linked from thematic portals in different countries. This topic will be addressed in a following section. This paper is divided in two parts and each one into sections. Part one is devoted to how data is generated and part two presents the current types of analysis that are performed and results that have been obtained. The comments at the end have three functions. The first is to ask some questions, the second is to try to answer them and, finally, the third is to describe the steps that are planned for the following months.

2. HOW DATA ARE OBTAINED

In order to understand how data are obtained, it is necessary to describe the technological environment of the ETD project – the Maxwell System.

2.1 The Maxwell System

The Maxwell System started in 1995 as a digital library to host learning objects and some communication and administrative tools for faculty and students to use. Learning objects were mainly class notes in both html and pdf, exercises, laboratory guides for experiments, etc. As usage became larger, the digital library was extended to support other types of digital contents that are found in a university – administrative documents, technical specs and manuals, simulators, etc. An extension of the distance learning and administrative functions became necessary when the system started being the platform for many distance learning courses. The recommendations of IMS – Instructional Management System Project ("IMS"), made available in 1998-2000, were used to develop version 3.0, currently in use.

In 1999, the development of the ETD module began and in 2000 the first ETD was published. In 2001, the incubation of UNICAP – Universidade Católica de Pernambuco ETD project started. Currently (Jun 15, 2005), there are 1,707 ETDs on the system (1,681 from PUC-Rio and 26 from UNICAP). In Aug 2002, ETDs became a requirement at PUC-Rio. Before that date, only some graduate programs required them. In 2002, the digital library was adapted to host senior projects and in 2003 the first online journal was published. Currently, there are 3 online journals and 2 more are due soon. The system is an institutional repository and has been one since it started supporting other types of digital contents besides learning objects. The expression institutional repository was not in use then. This characteristic of the system is important because it impacts the way data are mined and stored for statistical purposes.

In terms of ETDs, the system uses MTD-Br – Padrão Brasileiro de Metadados de Teses e Dissertações (Brazilian Standard for Metadata Elements of Theses and Dissertations) ("BDTD"). MTD-Br contains ETD-ms – An Interoperability Metadata Standard for Theses and Dissertations ("NDLTD"). The system is an OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting ("OAI") data provider. Its metadata are harvested by BDTD – Biblioteca Digital de Teses e Dissertações (Digital Library of Theses and Dissertations – the Brazilian national project) ("BDTD"). BDTD has a metadata union catalog of more than 12,000 full-text Brazilian ETDs. From BDTD, metadata are harvested and made available from NDLTD – Networked Digital Library of Theses and Dissertations ("NDLTD") and Chilean Cyberteses Net ("Cyber"). Some selected works are linked from thematic portals.

2.2 Mining Data from the Apache Server and Storing on th Data Base

The Apache Server log contains information of all operations that are performed with the system. In the last 6 months, the average number is 25,600 lines per day. But not all lines are associated to accesses to digital objects – there are clicks on buttons, browsing on Web pages, querying the database in search of contents, viewing statistics, accessing course schedules, looking for dates of exams, etc. When access statistics of digital contents came to discussion, it became clear that some definitions on how to extract data from the logs, i.e., how data should be mined, were necessary.

- 1st Definition: Visits and complete visits – The number of visits is the sum of all accesses to all digital objects of an ETD in a given month. A complete visit is a set of visits to all digital objects from a country in a given month.
- 2nd Definition: Country x IP address – The decision to use the country and not the IP address to establish a visit was based on the fact that the visits to an ETD can be made at different times (and reconnecting may assign a new IP address) and from different locations (with fixed IP addresses).
- 3rd Definition: Counting visits from the same IP address – Visits from the same IP are counted individually due to the fact that networks with many machines can be identified by the IP address of a firewall.
- 4th Definition: Counting visits to restricted digital objects – Some ETDs are totally or partially restricted – approximately 30% have some type of permanent or temporary restriction. Metadata, abstracts included, are publicly available for all of them. It was decided that attempts followed by a denial of access would be counted as accesses.

Once the definitions were made, the development started. The table was loaded and automatic update began in the middle of January 2005. The first statistics became available at the same time.

3. HOW DATA ARE COMBINED

After mining and storing data on the database, it is important to analyze and obtain results. This part presents the initial steps of the analysis and some results. The mined data and metadata on the database are used.

3.1 The Initial Combinations

As in the case of mining data, some definitions were necessary to start the analysis.

- 1st Definition: What lines to mine – Since the interest was on access to digital objects, the decision was to get the lines with extensions .dcr, .doc, .htm, .pdf, etc. All possible extensions on the database are considered, as long as the corresponding item is cataloged on the digital library, so that an eventual static html page is not counted.
- 2nd Definition: What information to get from the line – Statistics were planned on a monthly basis. The model treats data as sequences of points with discrete-time intervals (Luenberger, 1979) of a month. Past months data are unchanged and current month is updated according to the 3rd Definition. For this reason, the month and the year are extracted along with identification of the digital object and the country of the IP address of access. IPs are resolved using a plug-in called GeoIP Free that is available with AWStats ("AW"). A table was created on the database to store the mined data.
- 3rd Definition: Update of the database – In order to keep statistics updated and not to have a lengthy updating process, the choice was to read the lines every hour. Every hour, at the full hours (00:00, 01:00, etc), the incremental lines are mined. Accesses are summed for each month-year-digital object-country, so the table is not very big. In the last 6 months the average number of lines per month is 10,000.
- 4th Definition: When to start – Logs of the Apache Server had been saved since Jun 01, 2004. So, either this date was used or a later one, for example Jan 01, 2005. The decision was to use all the logs and this took some days of offline processing to load the table. From this event on, all loading is incremental and automatic. The (original) logs are stored and saved offline in case some change in the mining strategy is decided.

The system has 2 types of statistics – the ones related to the generation of digital objects and another set to examine access to them. The statistics of generation of digital contents have been available for many years and there are 15 only for ETDs, out of a total 27. The development of access statistics has been going on since January 2005. It is a long term project that will extend to other digital objects; there are 2 statistics for the online journals and bulletins. At the moment, there are 7 types of statistics for ETDs:

- Type 1: Author
- Type 2: Visited ETDs by month, year and country
- Type 3: Visited ETDs by country, month and year
- Type 4: 25 most visited ETDs
- Type 5: 20 most visited ETDs by institution
- Type 6: 10 most visited ETDs by graduate program / area
- Type 7: Graduate program / area

It is an objective of this project to relate accesses to countries of origin, to authors, contributors, graduate programs and areas. The current set of statistics is the embryo of a wider project whose next steps are already defined, but whose longer term strategy will be decided by studying current results.

3.2 The Initial Results

All ETDs in the collection are written in Portuguese and have titles, abstracts and keywords in at least one foreign language; the most common is English but French, Spanish and German are also used. Due to this fact, the initial results were quite surprising. Portuguese is the 3rd most spoken among the western languages (Ethnologue, 2005), just after Spanish and English, but most of its speakers are in Brazil. Some facts about the results:

- **Brazil accounts for about 55% of accessed ETDs**

The results of Type 1 statistics have been consolidated in 5 groups: Brazil, USA, Portuguese speaking countries (other than Brazil), Spanish speaking countries and others. Table 1 shows the consolidated results of accesses from June 01, 2004.

Table 1 – Accesses to ETDs from Jun 01, 2004

Group	Accesses (%)
Brazil	55.18
USA	12.18
Portuguese speaking countries	6.57
Spanish speaking countries	12.29
Others	13.78

Considering the information in table 1, it is easy to see that:

- Portuguese speaking countries account for 61.75% of the visited ETDs;
- Spanish speaking countries contributed with 12.29% – Portuguese and Spanish are very similar languages and it is not hard for speakers of one to read the other;
- The USA contributed with 12.18% of the visits – there is a large Spanish speaking group in the country.

It is reasonable to conclude that this result is language related.

- **Top 10 countries**

Since the logs started being collected, 114 IP locations have visited ETDs – 112 countries, IP = unidentified country and a2 = Satellite Access Host. Table 1 shows the top 10 countries in terms of visited ETDs in this time frame. The top 10 countries in terms of visits (from Jun 01, 2004 to Jun 15, 2005) are: Brazil (9,073), USA (2,002), Portugal (1,007), Spain (544), Peru (456), Mexico (277), Chile (210), France (167), Colombia (133) and Argentina (118). The relation of languages and geography is clear when table 1 and the numbers above are compared. France is the only country which is neither in the Americas nor in the Iberian Peninsula, but speaks a Latin language.

- **Different behaviors of 'best-sellers'**

Some ETDs have significant numbers of accesses every month from many different countries and this happens from the moment they are available on. This characteristic is found in the 3 centers the university is divided into. At the same time, there are some ETDs that have many visits just after they are published (approximately for a month or a month and a half) and then they receive a lot less visits. At third type is of ETDs that have few visits in the beginning, then, after some time, the numbers of visits grow. Since data have been available for 1 year only and analysis started a few months ago, it is expected that other types of behaviors will appear.

4. COMMENTS

The last part of this work can not be called conclusion because there are more questions to be asked than good answers for them. Some possible reasons for the results are suggested for further research and discussion.

- **Availability of ETDs in Portuguese and Spanish on NDLTD union catalog**

Examining NDLTD union catalog, ETDs in Iberian languages (Catalan, Portuguese or Spanish) were predominately from Brazil. Table 2 shows this result on Jun 15, 2005.

Table 2 – ETDs in Iberian languages on NDLTD union catalog on Jun 15, 2005

Institution	Country	Language(s)	Number
National Library	Portugal	Portuguese	185(1)
IBICT (includes PUC-Rio)	Brazil	Portuguese(2)	11,118
Various institutions	Spain (Catalunya)	Catalan or English or Spanish	2,066

(1) Most of them are not full text, only the initial pages; (2) About 200 in other languages.

The number of Brazilian ETDs is 4 times as big as the sum of all others. It is about 80% of the total offer.

Question: Can it be true that users of NDLTD union catalog who want works in an Iberian language are directed to Brazilian ETDs because of the offer?
Answer: It seems reasonable to think that YES.

° Uniqueness of the works

A reason for a big number of accesses is, no doubt, the uniqueness of some topics. There are 2 very important examples to be presented. The first is the area of Social History of Culture in the Graduate Program in History. There is a big focus on Brazil, specially in architecture at the turn of the century (19th – 20th) and the first half of 20th. Their T&D are well known in many countries and there is a link from the site of the Library of the Hamburg University of Technology [15], Germany, to some ETDs in the History Program. The second is the area of Children Education in the Graduate Program of Education. There is a big focus in Brazilian and, specially, in Rio's problems and characteristics. The results of Type 4 statistics have been consolidated. It was shown that for the year starting June 01, 2004, the numbers are:

« Social History of Culture – 2.3% of the available ETDs and 19.5% of the ETDs in the most visited list;
« Children Education – 3.4% of the available ETDs and 13.8% of the ETDs in the most visited list.

Question: Is the uniqueness of some subjects the reason of some many visits?
Answer: It seems reasonable to think that YES.

° Last published ETDs

The system has 2 functions that help users access the last works that are published.

« The last 10 published on the system;
« The last 10 published from a graduate program.

The published ETDs monthly average, starting Jun 2004, is 46.67. This means that an ETD remains among the 10 last visited roughly between 5 and 6 days.

At the same time, the numbers are very different if the second function is considered. Some graduate programs have many ETDs per year – Electrical Engineering had 50 in 2003 and 56 in 2004. On the other hand, other programs have few – Chemistry had 11 and 15, respectively, in 2003 and 2004. This means that Electrical Engineering ETDs remain among the last published less time than the ones from Chemistry. For example, on Apr 30, the oldest in the Electrical Engineering on last published list dated Feb 02, while in the Chemistry list the date was Dec 17, 2004.

Question: Can the time in the last published lists be a reason for some ETDs to have many visits in the beginning and then fade away?
Answer: This is a point to be investigated.

- **Links from other sites**

Some of the 'best-sellers' are linked from specific sites on the topics they address.

Question: Can this be significant in terms of the accesses?
Answer: This is a point to be investigated with the questionnaire that is mentioned in the following section.

- **Next steps**

The project of creating and analyzing statistics is meant to continue and to be extended to other digital contents. Preliminary results yield many interesting points, indicating that attention should be given to further analysis.One important information to gather about accesses to ETDs is how users got to the system – Google, NDLTD union catalog, BDTD union catalog, etc. To address this problem, an online questionnaire will be made available for users to inform administrators. This can be complemented by contacting advisors to check if ETDs are used as references to courses. It is interesting to remark that ETDs are specialized and require a high level of education to be read.

As this project progresses, further analysis can be carried out to separate areas or graduate programs that are 'more international', i.e., are visited by a larger number of countries, or that are visited from specific countries or regions.

5. REFERENCES

Awstats: http://awstats.sourceforge.net/

BDTD – Biblioteca Digital de Teses e Dissertações: http://bdtd.ibict.br/bdtd/

CyberTeses Net: http://www.cyberteses.net/

ETD-ms – An Interoperability Metadata Standard for Theses and Dissertations: http://www.ndltd.org/

Ethnologue – Languages of the World: http://www.ethnologue.com/

IMS Global Learning Consortium, Inc.: http://www.imsproject.org/ , http://www.imsglobal.org/

Luenberger, David G (1979), *Introduction to dynamic systems*, John Wiley, USA

Maxwell System: http://www.maxwell.lambda.ele.puc-rio.br

MTD-Br – Padrão Brasileiro de Metadados de Teses e Dissertações: http://bdtd.ibict.br/bdtd

NDLTD – Networked Digital Library of Theses and Dissertations: http://www.ndltd.org/

TUHH Universitätsbibliothek: http://www.tu-harburg.de/b/kuehn/