Building Data Warehouse

Teh Ying Wah, Ng Hooi Peng, and Ching Sue Hok

Department of Information Science
University Malaya
Malaysia
E-mail: tehyw@um.edu.my

Abstract

This paper introduces a framework for building the library data warehouse. It describes the steps in development of library data warehouse especially the extracting of data, transforming the data and loading the data into database. Much of the time was spent in these tasks due to the complexity of data. As the data may come from various sources, the amount of time spent on these tasks is often underestimated. In order to reduce the time consumed, we come out a systematic process of crawl only the data that we need and insert the data into database, instead of simply crawling all the data without planning and organise the data structure. As the data may come from various data source, issues on how we grab the data from various data source and store it into database will be discussed further. There are other considerations such as while extracting the data, there is a possibility of power failure and how we resume the crawling process and also how we grab the only data that is needed and also how we deal with the link or URL that is no longer available in the crawling process. The building of a data warehouse for library is an iterative process as the library data warehouse will be growing and evolving. Hence, flexibility and extendable issues are important as our framework will include this portable feature.

Keywords: library data warehouse, crawling, frameworks, eLibrary

1. Introduction

Data mining is no longer a new term in information science. Many researches have been done especially data mining in library which is also called bibliomining (Nicholson, 2003). The reasons of that include library has a huge collections of information as to provide these useful resources to library communities. Furthermore, the data are growing which may causes the size of database become Terabyte or even Petabyte. Moreover, valuable information might be hidden in this huge database and also due to the advanced technology recently, it has enable the possibility of data mining in library. As a result, systematic efforts are needed in order to develop a data warehouse for easing the use of data mining techniques.

The goal of this paper is to explore steps in developing a data warehouse for library and hence provide a framework that simplifies the process of developing a library data warehouse. First, the major steps of building a data warehouse will be explored. The steps included extraction, transformation and loading the data into database. Based on this concept, we developed our own library data warehouse and the process that we involved will be presented in this paper.

2. Related work

Building a data warehouse sounds simple. There is a temptation to think that building a data warehouse is only extracting the operational data and entering it into the data warehouse (Inmon, 2002). However, creating a data warehouse is more than that. Extracting data from various data source is a complex problem. Many inconsistency issues need to be deal with while integrating the data. Another issue that we need to look into is the efficiency of accessing the existing system data. Due to these complexity issues, much Extract-Transform-Load (ETL) software is available nowadays in easing the enterprise to build data warehouse. This software had improved over the traditional methods which extract using the SQL command manually and provide a tool helping them in creating ETL process. These ETL tools will be chosen according to the requirement. In other word, the ETL tools are customized to provide the functionality to meet the enterprise requirements. Hence, many of them choose to build their own data warehouse themselves. However, this approach is a risky business as many of the companies that choose to build data warehouse themselves were dissatisfied with the results although they had spent millions of dollars on it. To reduce the risk, framework for building a data warehouse can be used. This approach involves less time, cost, and risk than building your own data warehouse (Guerra et al., 2007). A framework can be customized to meet certain requirements. Therefore, obtaining a framework is the safety approach as it will guide you through out the building process compare to build the data warehouse from a scratch. Hence, we have come out our own framework for building the library data warehouse.

3. The Major Steps in Developing Data Warehouse

3.1 Identify the data source

The very first step before you start to develop data warehouse, the data source will be identified. You need to figure out what are the data that are required to be put into your data warehouse.

For a library data warehouse, there are two types of data sources that need to be considered, internal and external data source. Internal data source will be the data that already exist in the library system. The external data source is the data that does not exist within library system (Nicholson, 2003). In this paper, we will focus on external data source.

3.2 Build customized ETL tool

Each data warehouse has the different requirements. Therefore, a customized ETL tool is the better solution in order to fulfill the requirements. For the library data warehouse, we choose our own extract program. We deal the inconsistency issues with our own transformation method and finally we load the data into the data warehouse database.

3.3 Extraction

This can be the most time consuming part where you need to grab the data from various data source and store it into the staging database. Much of the time and effort are needed in writing a custom program to transfer the data from sources into staging database. As a result, during extraction, we need to determine which database system will be used for the staging area and also figure out what are the necessary data that are needed before grab it. The decline in the cost of hardware and storage has overcome the issues on avoiding the data duplication and also their worries on lack of storage as storing the excessive or unnecessary data. However, there is probably no reason to store the unnecessary data which had been identified not being useful in decision making process. Therefore, there is a necessary for extract only the relevant data before bringing into data warehouse (Mallach, 2000).

3.4 Transformation

After extracting the data from various data sources, transformation is needed to ensure the data consistency. In order to transform the data into data warehouse properly, you need to figure out a way of mapping the external data sources fields to the data warehouse fields. Transformation can be performed during data extraction or while loading the data into data warehouse. This integration can be a complex issue when the number of data sources getting bigger.

3.5 Loading

Once the extracting process, transforming and cleansing has been done, the data are loaded into the data warehouse. The loading of data can be categorised into two types; the loading of data that currently contain in the operational database and the loading of the updates to the data warehouse from the changes that have occurred in the operational database. As to guarantee the freshness of data, data warehouse is needed to be refreshed to update its data. Many issues are needed to be considered especially during loading the updates to the data warehouse. While updating the data warehouse, we need to ensure that no data are loosed and also to ensure a minimum overhead over the scanning existing file process.

4. Building the library data warehouse

A data warehouse is not enough just to have the internal data source and usually valuable data are from the external data source in decision making as per data warehouse main purpose. Internal data sources are easy integrated to data warehouse compare to external data sources where the complexity and unpredictable customise ETL process highly needed. External data sources typically and easily obtain is via internet web resource in html format.

Basically external data source major step involved in integration of data to data warehouse as discussed above which are the below 4 steps as shown in Figure 1:

1. Data Source Identification
2. Data Extraction
3. Data Transformation
4. Data Loading

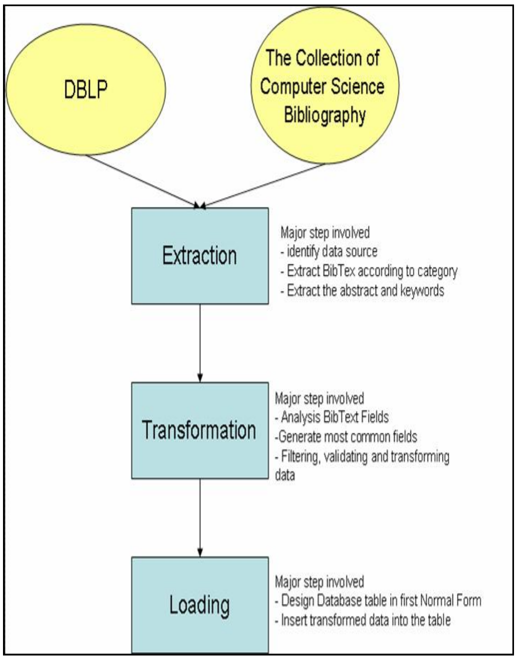

DBLP
The Collection of Computer Science Bibliography
Extraction
Major step involved
- identify data source
- Extract BibTex according to category
- Extract the abstract and keywords
Transformation
Major step involved
- Analysis BibText Fields
-Generate most common fields
- Filtering, validating and transforming data
Loading
Major step involved
- Design Database table in first Normal Form
- Insert transformed data into the table

Figure 1: Major step involved in integration of data

4.1 Data Source Identification

Some major computer science bibliography has been chosen as our external data source. The external data sources include **Digital Bibliography & Library Project** (DBLP) and **The Collection of Computer Science Bibliography**. These are E-libraries which provide large collection of resources in an index form. We crawl all information about computer science journals and proceedings from these data sources.

4.2 Data Extraction

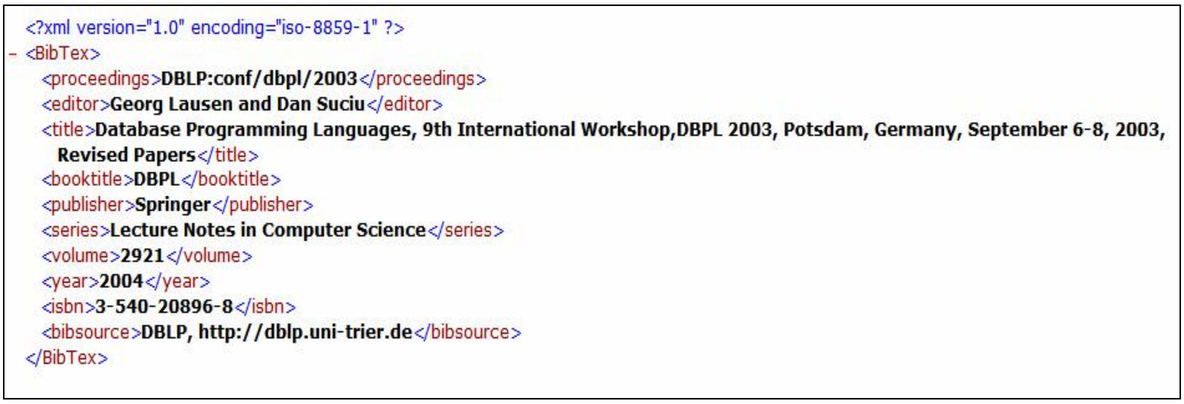
After identify the external data source, we figure out the data fields and start to crawl the data. From the source, we found that they have BibTex file format. BibTex is probably the most common format for bibliographies on the Internet. It has the standard entry types and standard fields. Hence, it helps during the transformation part; integrating the various data sources. As a result, we choose to crawl the BibTex files.

Web crawler or web spider is a program used for identified the links with BibTex information in the external web resource. Web spider is needed to be configure to crawl links with BibTex information, in our case, identified all the BibTex information is under a link path pattern for example http://.../rec/BibTex/* . The extracting program that we had chosen is open source, where spider.

We also crawl the abstract and keywords of the papers as to produce an informative library data warehouse. The abstract and keywords can be crawled from the link provided in the BibTex.

4.3 Data Transformation

When integrating different data source, a flexible data model is needed. XML provide this kind of flexibility therefore we get all the links and then download all the BibTex information and convert it to xml format as shown in Figure 2 for better manipulation of the data before transforms into data warehouse. By using XML, it helps in simplifying the data transformation part. All the data will be filtering, validating and transforming for ensure us getting the correct and desire data.



```
 <?xml version="1.0" encoding="iso-8859-1" ?>
- <BibTex>
   <proceedings>DBLP:conf/dbpl/2003</proceedings>
   <editor>Georg Lausen and Dan Suciu</editor>
   <title>Database Programming Languages, 9th International Workshop,DBPL 2003, Potsdam, Germany, September 6-8, 2003,
     Revised Papers</title>
   <booktitle>DBPL</booktitle>
   <publisher>Springer</publisher>
   <series>Lecture Notes in Computer Science</series>
   <volume>2921</volume>
   <year>2004</year>
   <isbn>3-540-20896-8</isbn>
   <bibsource>DBLP, http://dblp.uni-trier.de</bibsource>
 </BibTex>
```

Figure 2: BibTex in XML format

4.4 Data Loading

In data loading, we design the table in first normal form and partitioning which has better performance in SQL selection and updating. Besides that, this can avoid overhead of join operation and tolerance for data inconsistency before integrate to data warehouse. Another ETL process performed to obtain abstract information (abstract of a book, journal or article) and keywords for increase the data quality. In every ETL process, recovery feature is vital in case whenever any failure occurs. A last entry log file can be created to track where the failure was. This will enable the program resume at the failure point instead of scanning again to track the failure and avoid data duplication.

5. Result

The main result will be a library data warehouse that is approaching to completeness in terms of full of computer science journal; article and papers as the data we grab is accumulated from different data sources. These are the data that we have collected so far.

5.1 The current bibliographic statistics are as below

Types	Total
article	196128
incollection	2069
inproceedings	447268
masterthesis	5
phdthesis	67
proceedings	365836
book	2035
www	8

Fields	Total
path	1013416
editor	282005
author	643499
title	1013416
journal	196029
booktitle	798242
publisher	359348
series	204454
volume	397421
number	181669
Year	1013416
pages	611570
isbn	319496
bibsource	1013416
abstract	37261

6. Discussion

Lots of powerful ETL tools are available nowadays, for example, the popular open source ETL, Kettle. Our framework is different compare to these available ETL tools in extracting part. Our extraction is from external source to database instead of database to database.

During data extraction, there are certain papers do not provide keywords. These keywords are important and meaningful as to support the searching keywords function for library users. Hence, we develop some data mining technique in order to generate the possible keywords. The keywords that we generate will be based on the title and abstract of the paper.

7. Conclusion

The goal of this paper is to produce a framework that simplifies the process of building a library data warehouse and sharing the knowledge and problems that we had facing during the development. Once the crawling program and cleansing program has been done, what we left is only the minor maintenance as long as the library remains the same system. As a library data warehouse needs to be updated, it requires an iterative process of building it. Through this iterative process, we need to enhance the crawling and cleansing process in order to achieve the consistency and guarantee an updated data warehouse.

8. Reference

Inmon,W.H. (2002). *Building the Data Warehouse,* 3rd Edition. New York: John Wiley & Sons.

Guerra,J., McGinnis,J. and Andrews,D. (2007). “Why you Need a Data Warehouse”. *A Special Report for JD Edwards Customers*. Retrieved June 30, 2007 from http://www.rapiddecision.net/downloads/Why%20a%20Data%20Warehouse%2007.pdf

Nicholson, S. (2003). “The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making”. *Information Technology and Libraries* 22(4).

Mallach, Efrem G. (2000). *Decision Support and Data Warehouse Systems*, United States: McGraw-Hill.