A Mathematical Model For Treatment Selection Literature

G. Duncan and W. W. Koczkodaj

Abstract Business Intelligence tools and techniques, when applied to a data store of bibliographical references, can provide a researcher with valuable information and metrics. In contrast to specialized research platforms that provide a number of analysis tools, such as the Web of Knowledge(TM) or PubMed(TM), the techniques discussed in this paper provide a more generalized approach that can be used with most bibliographical datasets as well as with a number of different analysis tools. As a point of reference, the system utilizes the Web Of Knowledge's (WOK) Web of Science (WOS) database schema, chosen because it provides a comprehensive number of bibliographical information fields. This paper will discuss how to transform WOK formatted data into an online analytical processing (OLAP) cube as well as provide a few examples of using this technology to analyze bibliographical information.

1 Introduction

Business Intelligence (BI) ”is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information” [1]. Coined in 1958 by IBM researcher Hans Peter Luhn, Business Intelligence has since seen broad acceptance by the business world. BI provides a number of important applications in the enterprise, including measurement, analytics and report-

G. Duncan
Computer Science, Laurentian University, 935 Ramsey Lake Road, Sudbury, ON P3E 2C6, Canada
e-mail: gg_duncan@laurentian.ca

W.W. Koczkodaj
Computer Science, Laurentian University, 935 Ramsey Lake Road, Sudbury, ON P3E 2C6, Canada,
Tel.: 705-675-1151 ext.2311
e-mail: wkoczkodaj@cs.laurentian.ca

ing [1]. The core of BI is in utilizing large stores of data (referred to as data marts or data warehouses) to enable ad-hoc analysis, measurement metrics and data mining which can then be used in order to influence and guide business decisions. While BI has found a home in a number of business sectors such as in the financial and healthcare industries, using it to aid in the meta-analysis of reference information is virtually unknown. In general, BI can be applied to any relational dataset.

The sample implementation discussed within this paper is based on Microsoft's standard information management and software development offerings. The database used will be SQL 2008R2 and standard TSQL, the front-end system is implemented in C# under Microsoft Visual Studio 2012 and analysis/OLAP services will be provided by SQL Server Business Intelligence Development Studio 2008 deployed to a SQL Server Analysis Services server (SQL 2008R2). The choice of platform was made due to familiarity to the author.

Summary

The proposed system will be described in the following steps:

1. Extract and Transform the bibliographical data such it conforms to the Web Of Knowledge format.
2. Serialize the transformed data into a SQL database.
3. Transform the serialized SQL data into data dimensions and a fact table arranged in a star schema.
4. Define the OLAP cube's data dimensions' measures and their inter-relationships.
5. Create an OLAP cube from the data and deploy the cube to an analysis server.

Once the data is deployed, standard BI and OLAP tools (such as Microsoft Excel, Microsoft SQL Server Business Intelligence Development Studio or Tableau) can be used in order to analyze the cube. Manual analysis of the cube via the use of DMX (data mining extensions) or MDX (multidimensional expressions) will be considered out of scope in this document.

2 Extract and Transform

The Web of Knowledge allows the user to download search results into a variety of formats, as does PubMed. Once downloaded, this data may need to be transformed into an intermediate format so that it's compatible with the underlying database schema (or at least compatible with the tools being used to import the data into the database). In order to perform such a transform of the data, the example implementation utilises XML (extensible markup language), a standards-based markup language that defines rules for encoding information in a human and machine readable format [3], and XSLT (extensible style sheet language transformations) a language for transforming XML documents [4]. XML is an extremely well supported format; it is the format of choice for many internet data interchange protocols (see, for instance, [5]), which makes it particularly suitable as an intermediary format for extract and transform operations. See Figure 2 for an overview of the process.

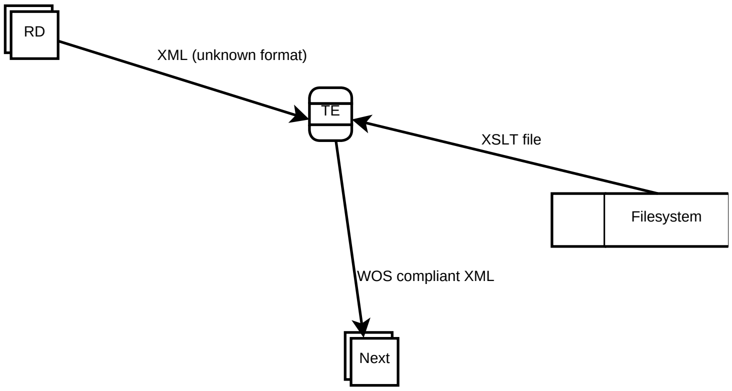

Fig. 1 Extract and Transform DFD. The user downloads XML from the reference database (RD) and sends it to the transform engine (TE). The appropriate XSLT files is loaded from the file system, and the transform is applied, creating an XML document compliant with the desired schema/format.

3 Serializing to a Database

Once the data is in the necessary format, the next step is to serialize it to a database. In the case of the example implementation, the data is placed into a SQL database table that mimics the schema of it's input data. This database will form the basis of the data warehouse from which the analysis services will query. It's not necessary that this database be normalized. See Figure 3 for a description of the process.

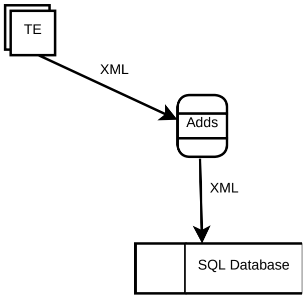

Fig. 2 Serializing the transformed data to a database. The transform engine (TE) sends the XML to the process that issues TSQL commands to insert the information into the SQL database.

4 Data Dimensions, Fact Table and the Star Schema

In order to create the OLAP cube, the data must be cut into dimensions and one or more fact tables must be created. The collection of dimensions, fact tables and data tables are collectively called the “data warehouse”. A data dimension is a “data set composed of individual, non-overlapping data elements”, the primary purpose of which is to provide “filtering, grouping and labeling” [2]. A fact table is a table that joins all the relationships between the various dimensions. The exact methods used to create the dimension and fact tables will vary depending on the a number of factors, such as the granularity of the fact tables and dimensions and the number and types of dimensions.

It is important to consider the data type and content, since some may not be very appropriate as dimensions. In some cases, the data may have to be broken up or otherwise manipulated in order to enable sufficient levels of granularity for effective analysis. For example, in Web Of Science formatted records, author and email information are each stored as semi-colon separated lists (“email1@domain.com; email2@domain.com”). In order for this field to be used as a data dimension, it is first necessary to separate out the individual values from their collated representations, and then individually insert them into the dimensional table. It is also important to consider how to key the dimensional information, (particularly the primary key, as this key will define one or more columns of the fact table). The typical technique used is to define the table’s primary key as an integer “identity” column. This provides for each row to be unique by definition. Other methods include the use of multi-part keys or object guids. Recall that the definition of a data dimension specifies that the data contained therein be non-overlapping and thus distinct.

Once the data dimensions have been defined and created, the next step is to create one or more fact tables from this data. The fact table provides 1 row for each valid combination of the dimension tables’ values.

The fact table is composed of all the foreign keys from each of the dimensional tables, along with any non-dimensional data from the basis table.

Figure 3 presents the entity relationship diagram created for the example implementation, notice it’s arrangement in the “star” pattern, where all dimensional tables are arranged around the central fact table.

4.1 *Fact and Dimension Table Choices in the Example Implementation*

For the example implementation, the following WOK record columns were chosen:

AU	Authors information
TI	Title
SO	Full source title
PD	Publication date

PY	Publication year
AB	Abstract
PT	Publication type
EM	Email addresses

These columns were chosen on the basis of their importance in terms of semantic content as well as the fact that they tended to be the most populated of all the WOK/WOS fields. Of special note are the AU, EM and AB fields.

Authors (AU) and Email Addresses (EM) Dimensions

Within the WOK/WOS schema (as mentioned previously), author and email address information is encoded into a semi-colon separated list. Thus given two authors for a paper A1, A2, the AU field would contain: “A1; A2”. In order to properly analyze this information, it must be decoded and each separate entity is inserted individually into the dimension table.

Abstract (AB) Dimension

In order to process the occurrence of certain words within the abstract, the entire abstract of each WOK record is treated as a collection of individual terms, separated by blank space, punctuation and parentheses. Using the same concept as with the Authors and Email addresses dimensions, each word and it’s associated WOK record are stored in the dimension table.

All together, the data warehouse for the example implementation is depicted in Figure 3. The number of rows in the fact table (which depends on the granularity of the dimensional tables), can grow to be quite large. For instance, a Web Of Science search that returned 296 rows, which, when properly dimensioned, produced a fact table consisting of 253,067 rows.

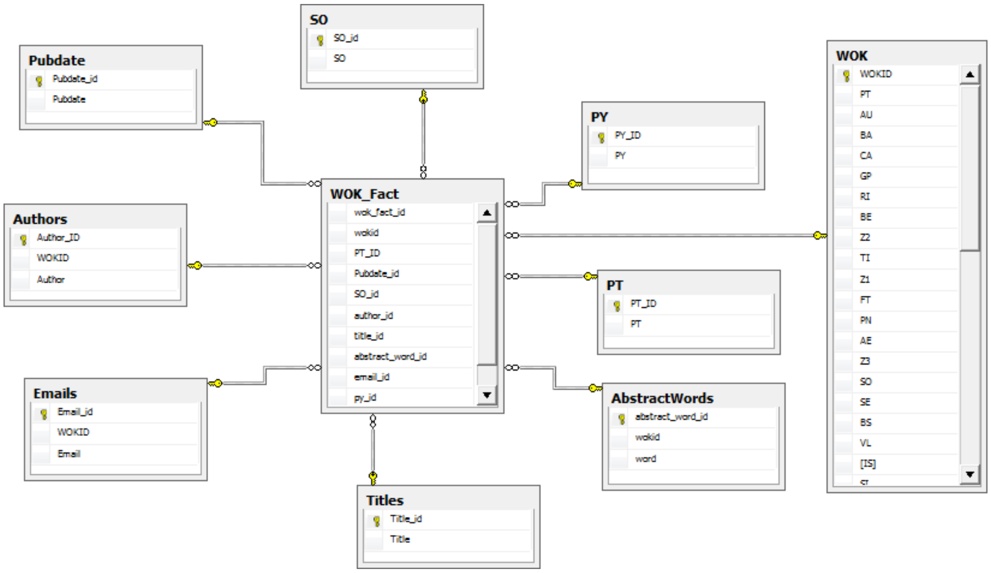

Fig. 3 Example implementation fact and dimension tables (data warehouse). Notice the "star" pattern of the layout.

5 Conclusion

Business Intelligence tools and techniques can greatly simplify the analysis of large amounts of data. By utilizing a common schema format defined in XML, the "WOS-ToDB" tool is able to serialize any conforming data to a SQL database. From this data is constructed an online analytical processing cube, which can then be used to quickly and efficiently analyze the data. While the implementation of a Business Intelligence project requires a certain amount of knowledge about data management tools and techniques, the end result is a re-usable system able to provide efficient and complex data analysis suitable to many varied problem domains.

References

1. Business intelligence. *Wikipedia, the free encyclopedia*, February 2013. Page Version ID: 540682028.
2. Dimension (data warehouse). *Wikipedia, the free encyclopedia*, February 2013. Page Version ID: 536574434.
3. XML. *Wikipedia, the free encyclopedia*, February 2013. Page Version ID: 539777964.
4. XSLT. *Wikipedia, the free encyclopedia*, February 2013. Page Version ID: 540812551.
5. Marshall T. Rose, Scott Hollenbeck, and Larry Masinter. Guidelines for the use of extensible markup language (XML) within IETF protocol. http://tools.ietf.org/html/rfc3470.