Intelligent Information Management, **2011, 3, 190-197**
doi:10.4236/iim.2011.35023 Published Online September 2011 (http://www.SciRP.org/journal/iim)

Warehousing and OLAP Analysis of Bibliographic Data*

Tsvetanka Georgieva-Trifonova
Department of Mathematics and Informatics, "St. Cyril and St. Methodius" University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria
E-mail: *cv.georgieva@uni-vt.bg*
Received June 25, 2011; *revised July* 4, 2011; *accepted July* 5, 2011

Abstract

In this paper, the system *bgMath*/*OLAP* for warehousing and online analytical processing bibliographic data is proposed. The implemented system can be useful for the users maintaining their electronic libraries with publications in order to monitoring, evaluating and comparing the scientific development of particular researchers, entire research groups, certain scientific fields and problems.

Keywords: Data Warehousing, Online Analytical Processing, Bibliographic Information System, Electronic Library

1. Introduction

In the last years, the digitalized form is important part of the creation, the distribution and the usage of the scientific literature. This fact concerns the periodical issues, as well as the conference proceedings and even the monographs and the reference books. The distribution and publishing the materials in Internet expedites the development of a large number of bibliographic systems integrated with search engines.

In [1], the system *Personal eLibrary bgMath* is represented, whose aim is different: easy manipulation with these materials (papers, dissertations, reports, etc.) and data about them, that are used repeatedly and are necessary at every turn in the scientific activity—scientific researching, writhing papers and dissertations, preparing reports and Web pages or application documentation. The bibliographic system for scientific literature *bgMath*, utilized from one or more users working at one or more scientific sections, allows accumulating data which are of interest for analyzing. This is the basic motivation for applying the warehousing and online analytical processing (OLAP) technology on the bibliographic data obtained from it.

The main purpose of the system *bgMath*/*OLAP* is to provide a possibility for monitoring, evaluating and comparing the scientific development of particular researchers, entire research groups, separate scientific areas and problems.

More concretely, the implemented system can be utilized for the following:

- Analyzing the bibliographic data collected from the usage of *bgMath* from one or group of researchers;
- Outputting the summarized reports about the number of the publications and the number of the citations of the publications of the particular authors, all members of departments, institutions by years and/or type of the publications;
- Monitoring the changes in the number of the publications and the citations in the different years.
- The basic features of the system *bgMath*/*OLAP* are divided by four groups:
- Loading the data in the data warehouse periodically by a given schedule;
- Calculating and maintaining the summarized data in the data cube;
- Browsing the summarized data with the purpose of their analyzing by different dimensions in a tabular and a graphical view through Microsoft Excel application;
- Exporting the summarized data in PDF, HTML, XML, others formats.

2. OLAP Systems and Bibliographic Databases

OLAP technology could be applied to support solving important problems regarding the bibliographic data-

*This research is partially supported by the project "Bibliographic system for organizing, storage and usage of the digitalized scientific literature", "St. Cyril and St. Methodius" University of Veliko Tarnovo, No RD 642-17/ 26.07.2010.

bases in the libraries. OLAP systems can be used for periodic reporting and data integrity checking. Analysts can interactively browse hierarchical and summarized data in order to extract new knowledge from the database. The traditional relational systems for database management that does not support OLAP, are appropriate for storing the data needed for daily activities and transactions processing. They are not suitable for performing complex queries that access large datasets and make multiple scans, joins and summaries, because they require more time for answer [2,3]. Minimizing the response time of these queries proves crucial influence at designing OLAP applications.

Recently OLAP systems on bibliographic databases are implemented. In [4], OLAP system for analyzing data in the Slovenian national bibliographic database *Biomedicina Slovenica* is proposed. DBPubs [5] is a system for analyzing and exploring the content of database publications by combining keyword search with OLAP operations.

In this paper, the system *bgMath*/*OLAP* is represented, whose purpose is applying OLAP technology to exploring the data obtained from the bibliographic system for scientific literature *bgMath*. The developed system allows analyzing the number of the publications and the citations by years, by keywords, by authors, by scientific sections, etc.

3. Designing and Implementing the System *bgMath*/*OLAP*

The development of the system *bgMath*/*OLAP* includes designing and implementing a data warehouse; a package for loading the data in the warehouse; a data cube; a client application for visualizing the results.

3.1. Architecture of the System *bgMath*/*OLAP*

The architecture of the system *bgMath*/*OLAP* is represented in **Figure 1**.

The relational database PublicationsDB is designed and created for the purposes of the system *Personal eLibrary bgMath*. The structure of this database is described in detail in [1].

The structure of the data warehouse PublicationsDW and the implementation of the process of data extraction, transformation and loading (ETL) are represented in Section 3.2.

On the basis of the dimension tables and the measures in the fact table defined in the design of the data warehouse, the dimensions and the measures of the OLAP data cube PublicationsCube are determined. The structure of the data cube is described in Section 3.3.1.

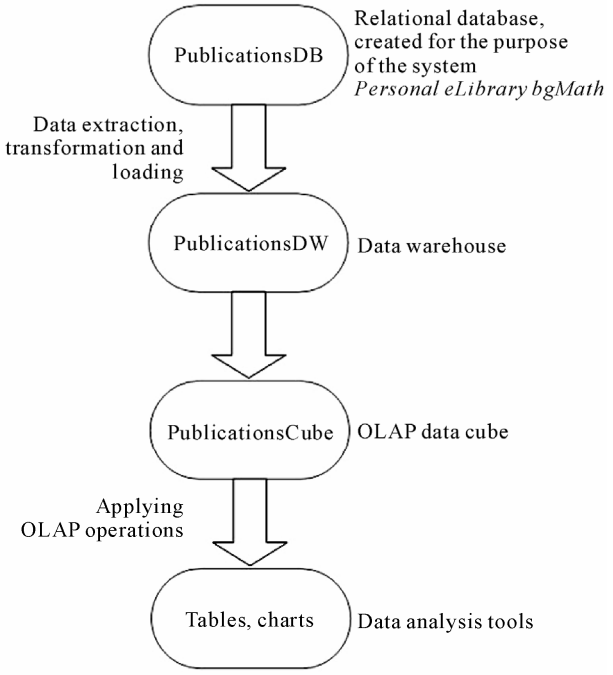


Figure 1. Architecture of the system *bgMath*/*OLAP*.

The usage of the OLAP data cube PublicationsCube for analyzing the bibliographic data is performed with an application that is implemented for the purpose with the means of Microsoft Excel. In Section 3.3.2, an exemplary table and a graph, which visualize summarized data from the data cube, are represented.

3.2. Warehousing the Bibliographic Data

The data warehouse serves for the data accumulation and organization with the aim of providing this data for analyzing. The purpose of the data warehouse determines the data model used for its designing.

3.2.1. Modeling Data in the Data Warehouse

The design of the data warehouses is based on the multidimensional model of the data [2,3]. This model includes several numeric measures, which are liable for analysis. Each measure depends on a set of dimensions. The normalization of the data is used for designing databases in OLTP (*Online Transaction Processing*) environment, but it is unsuitable for designing data warehouses. The physical implementation of the multidimensional model requires two types of tables: dimension tables and fact tables.

The multidimensional model can be represented with a star schema, a snowflake schema or a galaxy schema. The model of the data warehouse PublicationsDW is designed in conformity with the snowflake schema (**Figure 2**).

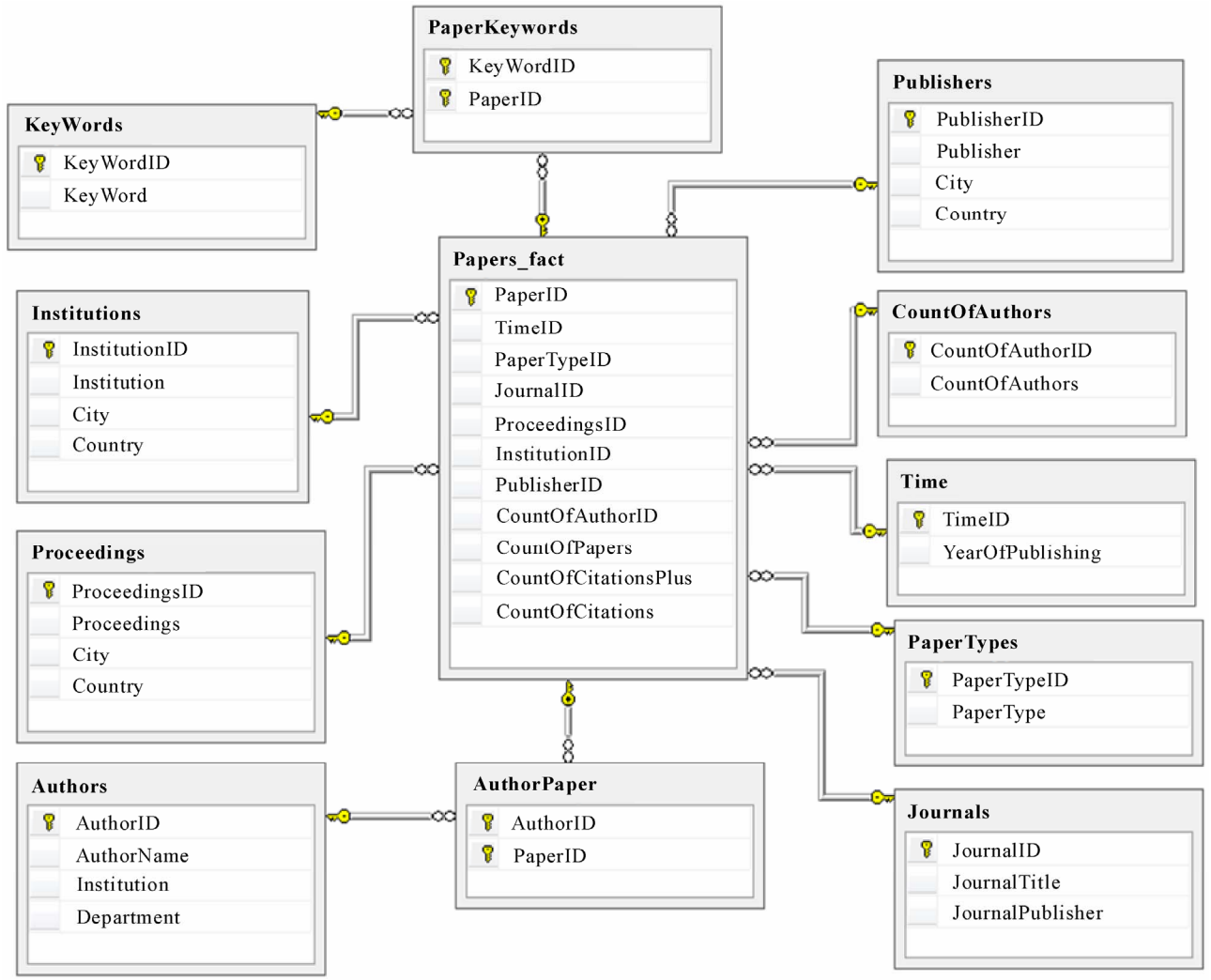


Figure 2. Snowflake schema of PublicationsDW.

The dimension tables in the data warehouse PublicationsDW store the data about authors; types of the publications; journals in which papers are published; conferences where papers are represented; keywords; year of publishing; number of the publication's authors. The fact table *Papers_fact* includes attributes which refer the dimension tables and the measure attributes: *CountOfPapers*—the number of the publications, *CountOfCitations Plus*—the number of the citations; *CountOfCitations*—the number of the citations without the self-citations.

We have taken advantage of the database management system MS SQL Server [6-12] to implement the data warehouse PublicationsDW.

3.2.2. Data Extraction, Transformation and Loading in Data Warehouse

The data loading into the data warehouse PublicationsDW is performed with a package created by using SQL Server Integration Services [13].

The following tasks are included in the package (**Figure 3**):

- Populating the dimension tables: *Authors*, *PaperTypes*, *Journals*, *Proceedings*, *Institutions*, *Publishers*, *KeyWords*, *Time*, *CountOfAutors*;
- Populating the fact table *Papers_fact*;
- Populating the tables *AuthorPaper* and *PaperKeywords*.

The service SQL Server Agent provides a possibility for creating a package job, which includes performing the package for data extraction, transformation and loading on given schedule.

Because the design of the data warehouse is based on the snowflake schema, the efficiency of the process of data extraction, transformation and loading in the warehouse is increased, consequently the efficiency of the system as a whole is improved.

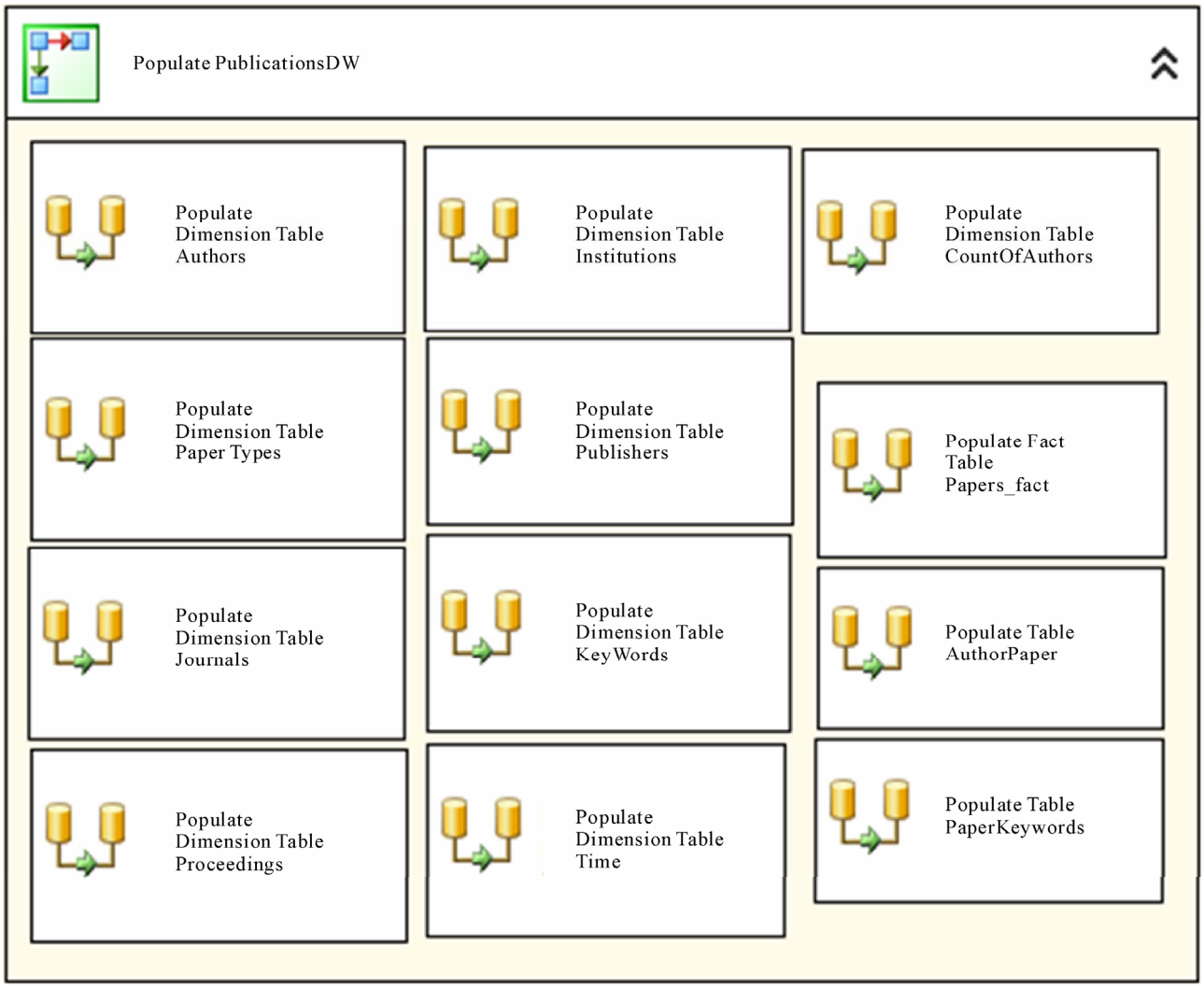


Figure 3. Loading data in PublicationsDW.

3.3. Online Analytical Processing the Bibliographic Data

Computing and sorting the summarized data, which are stored separately from the data sources for online transaction processing, decreases the quantity of the data for processing, when it is necessary for the users to analyze large amount of information. The organization of the data in the data warehouse into the structures corresponding to the multidimensional model and their previously processing provides maximal performance for the queries, which summarize the data by different ways.

3.3.1. Designing and Building the Data Cube

The data cube is a structure intended for providing fast access to the data in the data warehouse. It is a basic target for analytical processing the data. The data cube stores previously computed summaries of the data. The creation and the usage of the data cube eliminate the necessity from joining the tables and re-computing the values returned from the most frequent executed queries.

The dimensions and the measures in the data cube are determined by the dimension tables and the measures in the fact table in the data warehouse. In the data cube PublicationsCube the following measures are defined:

- The number of the publications—*CountOfPapers*;
- The number of the citations—*CountOfCitationsPlus*;
- The number of the citations without the self-citations—*CountOfCitations*.

The values of the measures are obtained in correspondence with nine dimensions:

- Publication's authors—*Authors*;

A hierarchy is defined for the dimension of the authors and the departments, the institutions where they work (**Figure 4**). Therefore the summarized data can be returned for chosen departments and/or institu-

tions.

- Publication's type—*PaperTypes*;
  The possible types of the publications are: books, journal's papers, conference's papers, dissertations, etc.
- Publication's journal—*Journals*;
  This dimension is also hierarchical. The data can be summarized for publishers of journals.
- Publication's conference—*Proceedings*;
- Publication's institute—*Institutions*;
- Publication's publisher—*Publishers*;
  The hierarchies are defined for the dimensions *Proceedings*, *Institutions* and *Publishers*. They allow performing summarization of the data by cities and countries.
- Publication's keywords—*KeyWords*;
- Year of publishing—*Time*;
- Number of the authors of the publications—*CountOfAutors*.

The structure of the data cube PublicationsCube created with SQL Server Analysis Services [14,12] is shown in **Figure 5**.

The data cube modeling includes defining hierarchies for some dimensions that has several advantages over creating separate dimensions:

- Improved usability—it is easy and comfortable for the users to work with the hierarchies that logically group the data;
- Shared aggregated data—if a dimension with a hierarchy is defined rather than building separate dimensions, the size and the complexity of the fact table is reduced.

Besides, the choice of the snowflake schema at data warehouse modeling leads to decreasing the redundancy of the stored data and the size of the obtained data cube.

3.3.2. OLAP Analyzing the Bibliographic Data

Microsoft proposes a standard query language for performing selection and manipulation of the multidimensional data—MDX (*MultiDimensional eXpressions*) [15, 12]. The result from an exemplary MDX query is shown

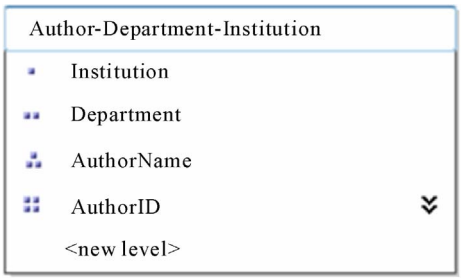


Figure 4. Hierarchy for the dimension *Authors*.

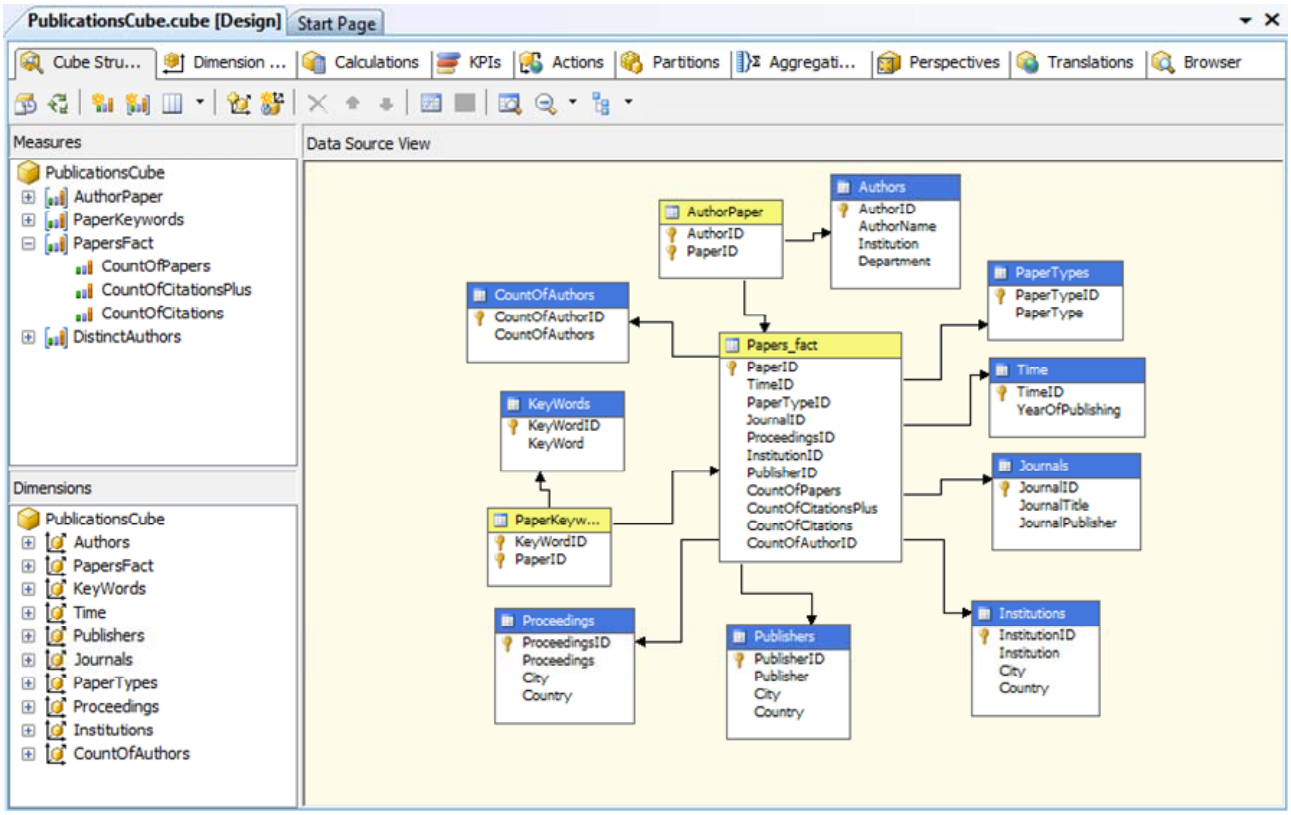
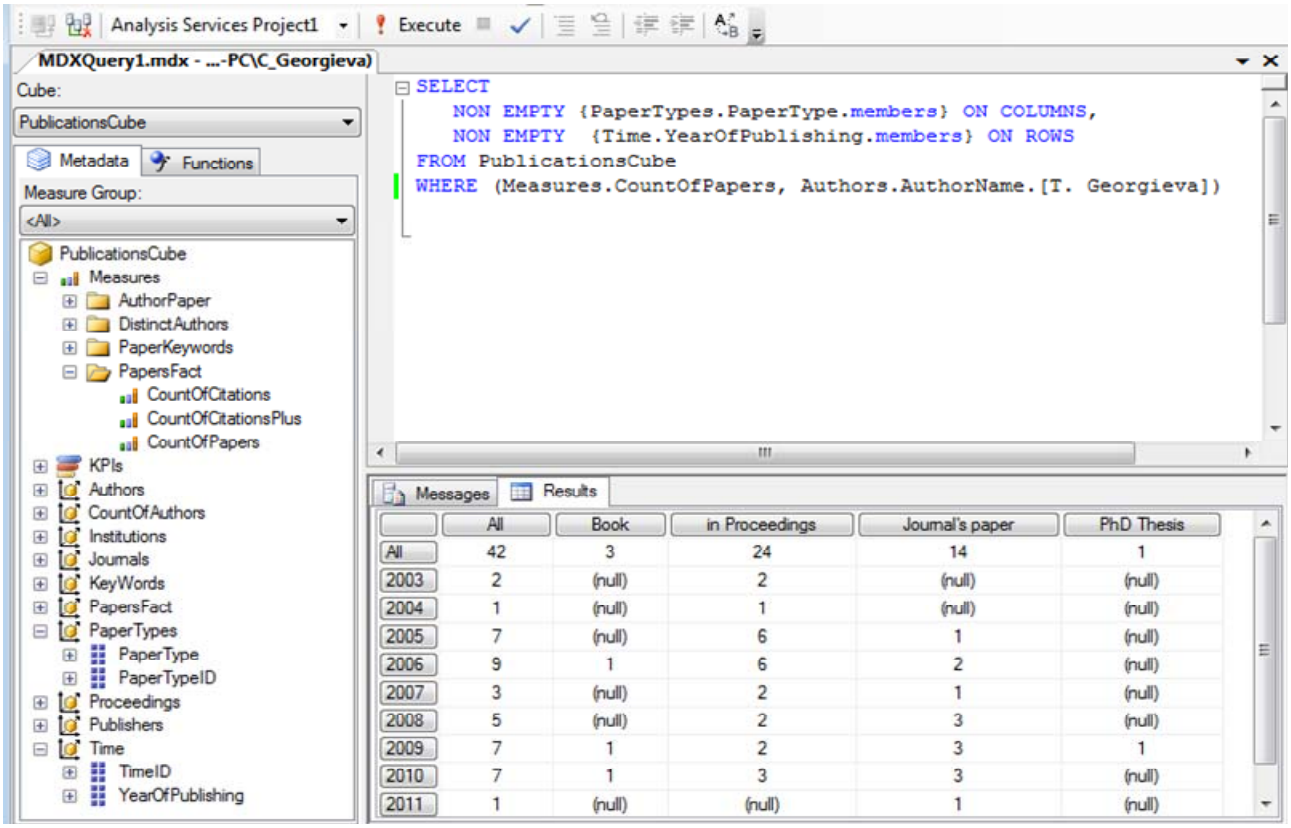


Figure 5. Data cube PublicationsCube.

in **Figure 6**. It is executed from the environment of Microsoft SQL Server Management Studio and extracts summarized data from the cube PublicationsCube. This query selects the number of the publications of a given author according to the types of the publications and the years of their publishing.

For the end user, an application is implemented with the means of Microsoft Excel [16,14]. This application allows extraction of the summarized data from the data cube PublicationsCube and their representation in tabular



	All	Book	in Proceedings	Journal's paper	PhD Thesis
All	42	3	24	14	1
2003	2	(null)	2	(null)	(null)
2004	1	(null)	1	(null)	(null)
2005	7	(null)	6	1	(null)
2006	9	1	6	2	(null)
2007	3	(null)	2	1	(null)
2008	5	(null)	2	3	(null)
2009	7	1	2	3	1
2010	7	1	3	3	(null)
2011	1	(null)	(null)	1	(null)

Figure 6. Execution of MDX queries from Microsoft SQL Server Management Studio.

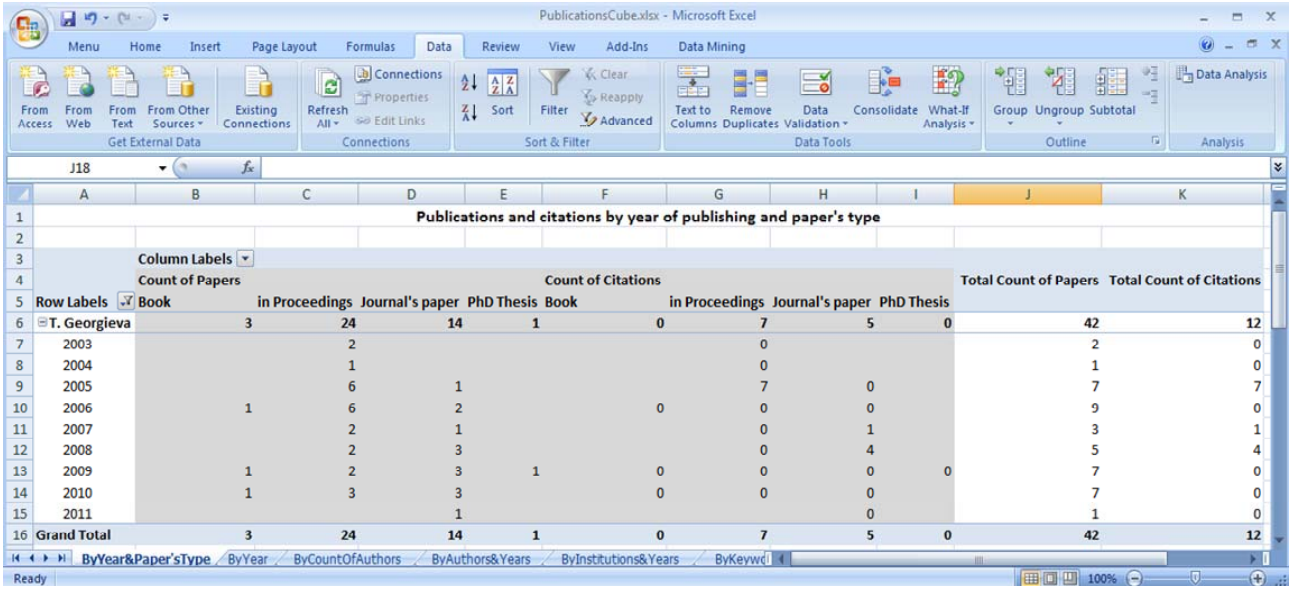


Figure 7. Analysis of the data through Microsoft Excel application.

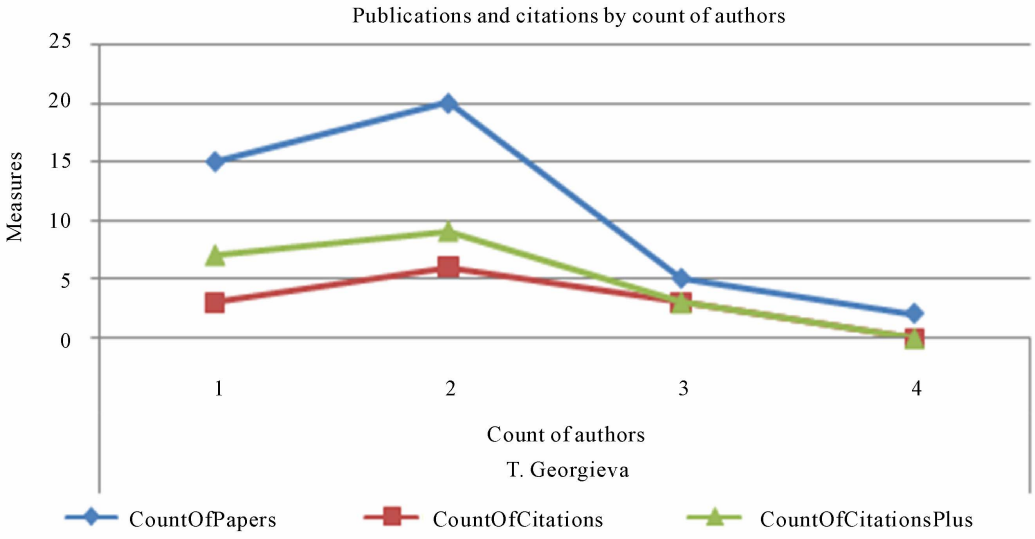


Figure 8. Analysis of the data through Microsoft Excel application.

and graphical views. The users can access multiple reports with the application and some of them are the number of:

- The publications and the citations by years and types of the publications (**Figure 7**);
- The publications and the citations by years;
- The publications and the citations by the number of the author of the publications (**Figure 8**);
- The publications and the citations by authors and/or years;
- The publications and the citations by institutions and/or years;
- The publications by keywords, authors and/or years;
- The publications and the citations by journals and/or years;
- The publications and the citations by conferences and/or years;
- The publications and the citations by publishers and/or years.

4. Conclusions

In the present paper, an application of OLAP technology for analyzing the bibliographic data is represented. The structure of the created data warehouse is described, as well as the implementation of the ETL process, the structure of the OLAP data cube, the features of the application designed for the end user.

Our future work includes applying the algorithms for data mining and development of an application providing possibilities for mining bibliographic data.

5. References

[1] I. Bouyukliev and T. Georgieva-Trifonova, "Development of a Personal Bibliographic Information System," *The Electronic Library*, 2011, accepted for publication.

[2] A. A. Barsegyan, M. S. Kupriyanov, V. V. Stepanenko and I. I. Holod, "Technologies for Data Analysis: Data Mining, Visual Mining, Text Mining, OLAP," BHV-Petersburg, Saint Petersburg, 2008.

[3] W. H. Inmon, "Building the Data Warehouse," Wiley Publishing, Inc., Hoboken, 2005.

[4] E. Hudomalj and G. Vidmar, "OLAP and Bibliographic Databases," *Scientometrics*, Vol. 58, No. 3, 2003, pp. 609-622. doi:10.1023/B:SCIE.0000006883.28709.d2

[5] A. Baid, A. Balmin, H. Hwang, E. Nijkamp, J. Rao, B. Reinwald, A. Simitsis, Y. Sismanis and F. Ham, "DBPubs: Multidimensional Exploration of Database Publications," *Proceedings of the 34th International Conference on Very Large Data Bases*, Vol. 1, No. 2, 2008, pp. 1456-1459.

[6] I. Ben-Gan, L. Kollar, D. Sarka and S. Kass, "Inside Microsoft® SQL Server® 2008: T-SQL Querying," Microsoft Press, Redmond, 2009.

[7] M. Coles, "Pro T-SQL 2008 Programmer's Guide," Apress, New York, 2008.

[8] L. Davidson, K. Kline, S. Klein and K. Windisch, "Pro SQL Server 2008 Relational Database Design and Implementation," Apress, New York, 2008.

[9] J. Mundy, W. Thornthwaite and R. Kimball, "The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset," John Wiley & Sons, Hoboken, 2006.

[10] P. Nielsen, M. White and U. Parui, "Microsoft® SQL Server® 2008 Bible," Wiley Publishing, Inc., Hoboken, 2009.

[11] V. Rainardi, "Building a Data Warehouse: With Examples in SQL Server," Apress, New York, 2007.

[12] G. Spofford, S. Harinath, C. Webb, D. H. Huang and F.

Civardi, “MDX Solutions, Second Edition: With Microsoft® SQL Server™ Analysis Services 2005 and Hyperion® Essbase,” Wiley Publishing, Inc., Hoboken, 2006.

[13] Microsoft Corporation, “SQL Server Integration Services,” 2008. http://msdn.microsoft.com/en-us/library/ms141026.aspx

[14] J. Shumate, “A Practical Guide to Microsoft OLAP Server,” Addison-Wesley, Boston, 2000.

[15] Microsoft Corporation, “Multidimensional Expressions (MDX) Reference”, 2008. http://msdn.microsoft.com/en-us/library/ms145506.aspx

[16] J. Krishnaswamy, “On Accessing Data from an OLAP Server Using MS Excel,” 2005. http://www.aspfree.com/c/a/MS-SQL-Server/On-Accessing-Data-From-An-OLAP-Server-Using-MS-Excel

