Application of Data Warehousing & Data Mining to Exploitation for Supporting the Planning of Higher Education System in Sri Lanka

M.G.N.A.S. Fernando and G.N. Wikramanayake

University of Colombo School of Computing
{nas, gnw}@ucsc.cmb.ac.lk

ABSTRACT

The topic of data warehousing encompasses architectures, algorithms, and tools for brining together selected data from multiple databases or other information sources into a single repository called a data warehouse, suitable for direct querying and analysis.

We have designed a data warehouse for the Sri Lanka education system and applied basic data mining techniques (i.e. data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge representation) to support decision making activities. For this we have built an integrated data warehouse consisting data from Dept. of Examination, University Grants Commission, School Census data, national population data and University student's information.

This paper highlights how the data warehouse was built for the Sri Lanka education system and how it was used to create data summary cubes for data analysis and mining process. At present using this developed system, basic level of summaries and analysis can be performed to obtain for decision support information. Further applying data mining techniques and advanced queries, we can obtain the necessary knowledge for policy marking as well.

1.0 INTRODUCTION

In Sri Lanka, there are limited opportunities to admission to the state universities. From 40% to 50% students who pass the G.C.E. (A/L) examination only 13% to 17% students can enter to the state university [8]. Among them although about 90% graduates annually [9], only 27% are able to find employment [5]. Thus the state universities contribute towards increasing the unemployment rate with 70% from the Arts Stream [5].

University admission policy was reviewed in 1984 & 1987 and some of the policies were implemented successfully while others are yet not implemented due to varies reasons [6, 7]. Although many changes have been introduced to the education system the graduate unemployment problem have still not been resolved. We feel the lack of integrated information for decision making as one of the core reasons. Policy decisions made years ago are being carried forward due to political pressures and lack of data to convince the need for such a change.

One of the contributing factors is the mismatch between the state and private sector requirements for employment. Many students seeking alternative non-higher education paths have been successful than some of the state graduates who have been ending up as unemployed. The present education policies and the university selection criteria could be contributing towards producing unemployed graduates.

This paper investigates how such issues could be addressed using basic data mining techniques. Our investigation is based on designing a data warehouse and applying data mining techniques to obtain information for decision making [10] and hence assist the policy makers. This process can be seen as in figure 1.

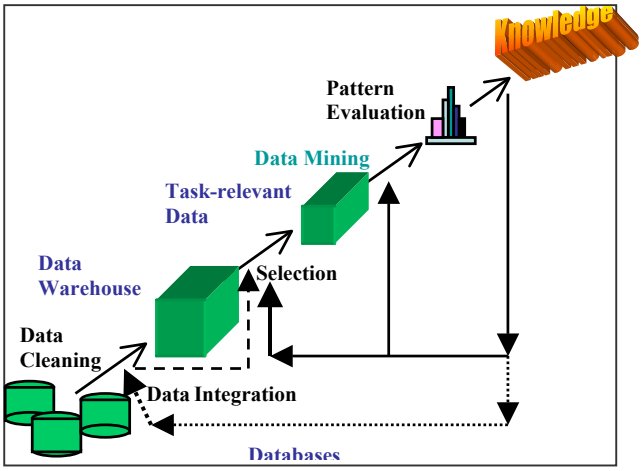

Figure 1: Data Mining Process

2.0 DATA WAREHOUSE

Data can now be stored in many different types of databases. One database architecture that has recently emerged is data warehouse, a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making [2, 3]. Data warehouse technology includes data cleaning, data integrating, and on-line analytical processing (OLAP) that is, analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information from different angles.

A data warehouse is a "subject-oriented, integrated, time variant, non-volatile collection of data that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions.

In data warehouses historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Query throughput and response times are more important than transaction throughput.

2.1 Multi-dimensional Data

To facilitate complex analyses and visualization, the data in a warehouse is typically modelled multi-dimensionally. For example, in a sales data warehouse, time of sale, sales district, salesperson, and product might be some of the dimensions of interest. Often, these dimensions are hierarchical, e.g. time of sale may be organized as a day-month-quarter-year hierarchy.

2.2 Data Analysis

Typically data analysis is done through OLAP operations such as rollup (increasing level of aggregation) and drill-down (decreasing level of aggregation or increasing details) along one or more dimension hierarchies, slice-and-dice (selection and projection) and pivot (re-orienting the multidimensional view of data).

2.3 Data Warehouse Architecture

Data warehouses are built using three-tier architecture as shown in figure 2. They are constructed via process of data cleaning, data transformation, data integration, data loading and periodic data refreshing. Data warehouse is a stored under a unified schema, and it usually resides at a single site. This is the bottom tier of the architecture and is managed by a data warehouse server.

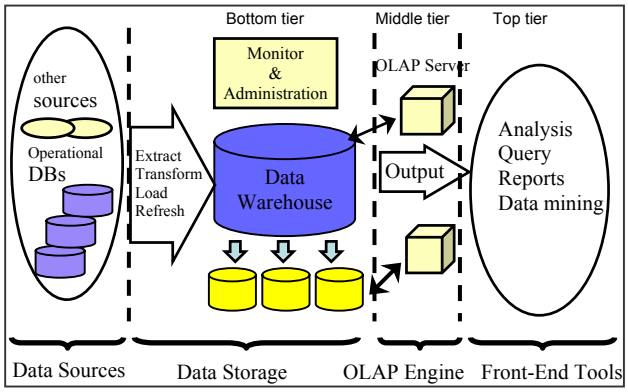

Figure 2: Three-tier Architecture

Analysing and query processing of huge data warehouse is very difficult and time-consuming task. Therefore multi dimensional data cubes consisting of summary tables are created for all possible decision marking activities [4]. The middle tier implements multidimensional data and operations and an OLAP server is typically used for that. The top tier is a client who uses front-end tools to obtain information.

2.4 Data Mining

Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. The spreadsheet is still the most compiling front-end application for OLAP. The challenges in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operation effectively over large multi-gigabytes databases. Indeed, the *essbase* product of Arbar cooperation uses Microsoft Excel as the front-end tools for its multidimensional engine [1].

3.0 EDUCATION DATA WAREHOUSE

3.1 Data Sources

The Education Ministry conducts annual school census through their Zonal Education Offices. They have data related to number of students in different streams of study for each school. The Dept. of Examinations conducts all national examinations and hence they have about the year 5 scholarship examination, G.C.E. (O/L) and G.C.E. (A/L) examinations. The University Grants Commission (UGC) process the university admissions and they have data related to the students admitted to each university for various streams of study. The university admission process is based on the national population and this data is maintained by the Dept. of Census. Each university has the academic performances of their students. However an important source of information of what the graduates are doing after completing the degree is not maintained by anybody,

except for time to time the government collects applications from unemployed graduates.

The way the data is maintained it is not possible to link this data to individuals such as how a person has performed from school time to university. However, characteristics identified at each stage could be linked to obtain valuable information such as industry acceptance for an average rural student.

3.2 Design the Data Warehouses

The data of the above sources has to be cleaned and transformed so that the different database structures could be integrated as shown in figure 3. Thereafter the data from the relevant data sources have to be loaded to the integrated data warehouse and later on periodically updated through import data from the respective data sources.

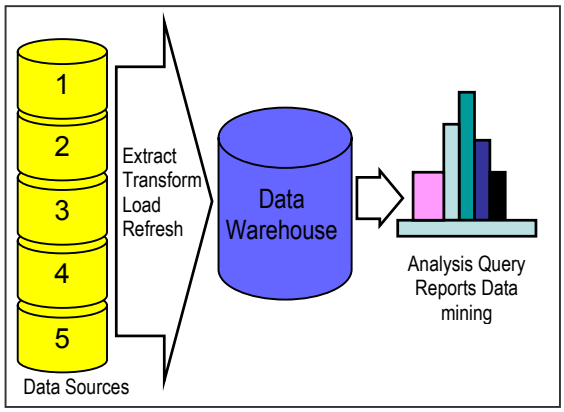

Figure 3: Education Data Warehouse

Data warehousing systems use a variety of data extraction and clearing tools, and refresh utilities for populating warehouses. Data extraction from different location is usually implemented gateways and standard interfaces (such as information builders SQL, ODBC, Oracle open connect Informix enterprise gateways). We used extraction tool available in SQL Server.

3.2.1 Data Cleaning

A data warehouse is used for decision marking. It is important that the data in the warehouse is accurate. As large volumes of data from multiple sources are involved there is a high probability of errors and anomalies in the data.

All our data sources had used internal controls to validate their data. Also data of the Dept. of Examination and UGC has to ensure 100% accuracy. Hence our data sources could consist of a little amount of data error and anyway there is no way of determining this.

Usually missing data are identified at this stage and necessary actions are taken. These data are either ignored, filled manually or automatically using a global constraint such as unknown or infinity, mean value or probable value. We came across missing values such as year, university etc. for some of the unemployed data. Presently we have opted for the ignore option to deal with such cases.

Identifying noisy data and smoothing them is also part of the data cleaning process. This may not apply for most educational data. Detecting and correcting data inconsistencies is also part of data cleaning.

3.2.2 Data Integration

Data integration is the next important step. All our data sources use the data mainly to produce annual summaries. Hence the data of different years are not checked for consistency. Also the data of different years need not have the same database structures. This makes the integration process over several years more difficult. To update the data warehouse, same format should be consider, therefore after modifying the data structures and after performing relevant processing we had to construct our data warehouse.

3.2.3 Data Selection

All the fields of the operational databases are not necessary for the data warehouse. Most important fields for the decision marking activities are selected and updated the warehouse. For an example All Island Ranks, District Ranks are available in the Department of Examinations data source and it will not update the data warehouse as these can be derived as well as the rank change if some does not apply. Details such as student details are not required for decision making.

3.2.4 Data Transformation

There were so many coding differences in the data sources. Codes used for district, school etc. by the different sources were quite different (e.g., in the census population, district code for Matale and Kandy were 4 and 5 respectively, while the Examination department and UGC had used Kandy and Matale as 4 and 5 respectively). Thus matching the structures only was not sufficient. Another common mismatch was the use of different data types and field lengths (e.g. actual size of the district code is two digits but some sources had declared it as a string and some sources it declared as number with 8-digit length). We used SPSS to convert the different data format into one form as well as recoding purposes.

After extracting, clearing and transformation data must be loaded into the warehouse. Additional pre-processing such as checking integrity constrains, sorting summarization, aggregation and other computation may

still be required to build the warehouse tables. Typically batch load utilities are used for this purpose. In addition to populating the warehouse, a load utility must allow the system administrator to monitor status so that in case of a failure operation such as cancel, suspend and resume a load, can be performed with no loss of integrity.

For our system data can be loaded yearly, after completion of their major operational works. For example, A/L databases can be loaded after updating A/L re-correction marks added, UGC selected students database can be updated after completing of filling of vacancies.

3.2.5 Refresh

Refreshing a warehouse consists of propagating updates on source data to corresponding updates in the warehouse. There are two sets of issues to consider: when to refresh and how to refresh, usually the warehouse is refreshed periodically (daily or weekly) only if some OLAP queries need current data, it is necessary to propagate every update. The warehouse administrator sets the refresh policy, depending on user needs or traffic, of the data sources. In this education decision support system all operational process are completed annually.

3.3 Star Schema

The most popular model for a data warehouse is the star schema. Here a fact table and set of dimension tables are identified. Multi-dimensional cubes are formed using these tables. The star schema for our education system is given in figure 4. It can be formally defined using the DMQL as in figure 5. The fact table is defined as a cube_master with the dimensions as the attributes (e.g. AL_Year) and facts (e.g. stu_status). Then each attribute is defined as a dimension (e.g. Year).

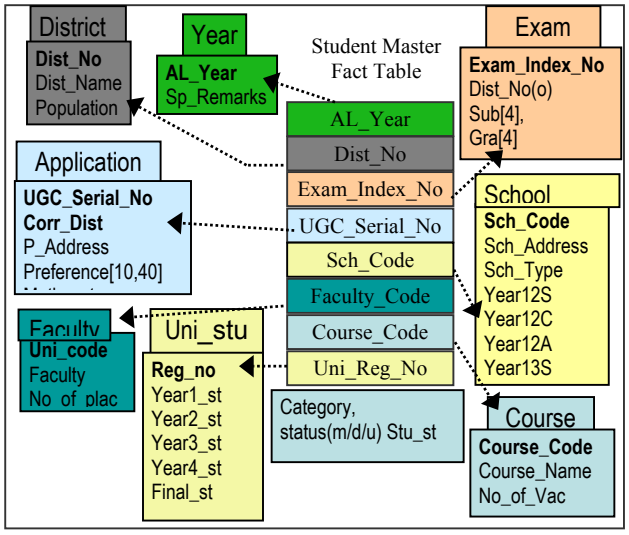

Figure 4: Star Schema

Just as relational query language like SQL can be used to specify relational queries. A data mining query language can be used to specify relational queries. We examine an SQL based data mining query language called DMQL, which contains language primitives for defining data warehouses. Data warehouses can be defined using two language primitives, one for cube definition and other one for dimension definitions.

```
define cube cube_master [AL_Year, Exam_Index_No,
      UGC_Serial_No, Uni_Reg_No, Faculty_Code,
      Course_Code, Sch_Code, Dist_No] : stu_status=
      count(*)

define  dimension  Year as  (AL_Year, Sp_Remarks)

define dimension District as (Dist_No, Dist_Name,
      Population)

define dimension Course as (Course_Code,
      Course_Name, No_of_Vac)

. . . . . .
```

Figure 5: part of DMQL statements

3.4 Performing Data Mining

Data mining represents the next quantum step beyond the historical and aggregate-based world of information that the data warehouse makes available to users. Data mining allows organizations to collect vital information regarding business processes, customers, sales and marketing and arrange the information in such a fashion as to allow business users to make predictive decisions about what direction the business should focus its resources. This advantage allows business decision makers to "steer" the focus of an organization and facilitate the continued success of the enterprise.

Educational data too can be mined to find hidden knowledge for the decision marking process. One of the major tasks is to review and examine the current policy for the admission to the higher education institutes, i.e. Present Quota System (Merit, District Under- Privilege and Special intake quota system), and to make recommendations if any changes are considered necessary etc.

In order to facilitate decision marking, the data in a data warehouse is organized around major subjects, such as year, district, course etc.. The data are stored to provide information from a historical perspective (such as the past 5- 10 years) and typically summarized. The data warehouse is usually modelled by a multidimensional database structure, where each dimensions corresponds to an attribute or a set of attributes in the schema and each cell stores the value of some aggregate measure.

4.0 DATA CUBES

There are so many billons of data records in a data warehouses. Analysis and query processing of such huge data warehouse is very difficult and time consuming task. Therefore maintaining multi-dimensional warehouse servers will help to solve this problem. Multi decisional data cubes consists summary tables for all possible decision marking activities. A possible data cube is shown in figure 6.

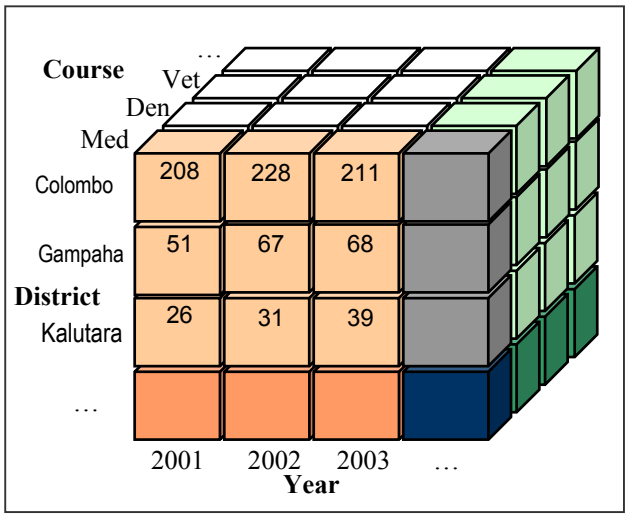

Figure 6: A Data Cube

After defining the star schema we can create so many cubs according to the requirements. For example as in figure 6, a cube may represent the selected students as a function of Year, District, and Courses (e.g. Year: 2001, 2002, 2003 etc., District: Colombo, Gampaha, Kalutara etc. and Courses: Medicine, Dental, Vet. Science etc. are the three dimensions with 208 students from Colombo district selected for Medicine in year 2001).

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to more general higher-level concepts. Using it data could be aggregated or disaggregated. Many concept hierarchies are implicit within the database schema and a hierarchy could be defined for the locations in the order of school < zone < district < province < country (see figure 7). This allows districts to be aggregated to provinces (roll-up) and well as districts to be disaggregated into zones (drill-down).

4. 1 Roll-up

The Roll-up operation corresponds to taking the current data objects and doing further grouping by one of the dimensions. The Roll-up operations performed on the central cube by climbing up the concept of hierarchy of figure 7.Thus, it is possible to Roll-up admission data by grouping districts into provinces. Hence Western Province would have 285 students selected for Medicine for the Year 2001 as in figure 8.

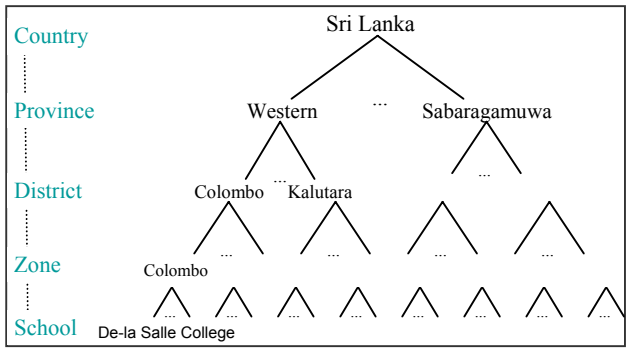

Figure 7: Concept hierarchy

Similarly for the location concept hierarchy the roll-up operation could be used to aggregate the school data to zone or zone to district or even province to country.

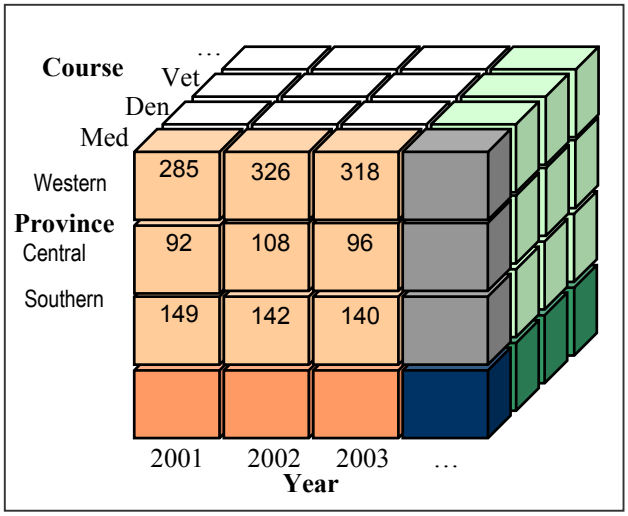

Figure 8: Roll-up District

4. 2 Drill-down

The drill-down operation is the opposite of roll-up. It navigates from less detail data to more details. The data cube of figure 6 can be drill-down using the location concept-hierarchy and hence districts can be disaggregated into zones as shown in figure 9.

4. 3 Slice and Dice

The slice operation performs a selection on one dimension of the given cube, resulting is a sub-cube. For example, we could select the course dimension and slice for course medicine and view a sub-cube as shown in figure 10.

The dice operations define a sub-cube by performing a selection of one or more dimensions. For example, three dimension dice for courses “Medicine” and Dental” for

districts “Colombo” and “Jaffna” for year 2001 and 2002 could be as shown in figure 11. Using this approach it is possible to focus on specific areas such as district or course and hence make appropriate decisions.

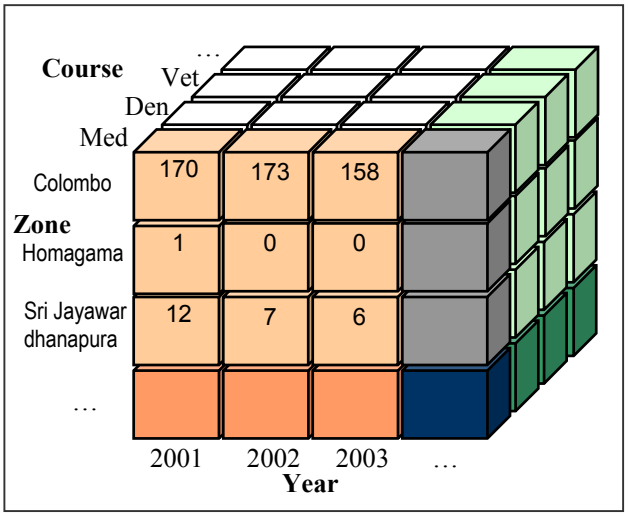

Figure 9: Drill-down District

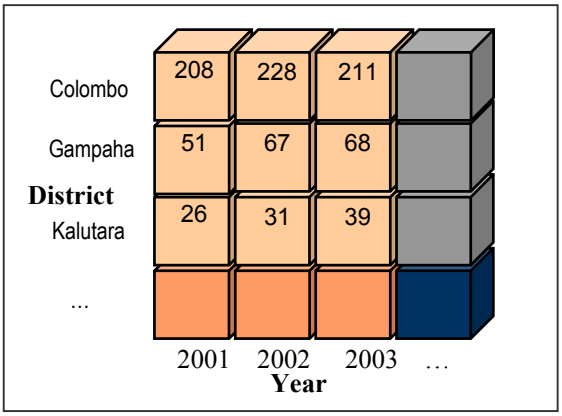

Figure 10: A Slice for Course Medicine

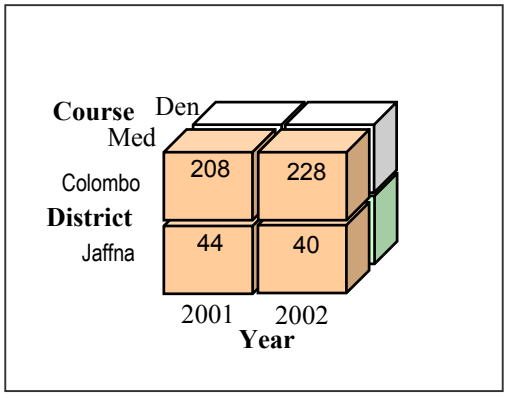

Figure 11: A 3D Dice

4. 4 Pivot (Rotate)

Pivot is a visualisation operation that rotates the data axes in view in order to provide an alternative presentation of the data. It allows visualisation of the other side of the dice.

5.0 CONCLUSIONS

For a developing country like Sri Lanka, education is more relevant factor for its entire economy. To contribute to the above, making correct decisions and implementing the correct policies are very essential. To achieve the above task it is important to integrate the heterogeneous data sources and use them to build a data warehouse and data cubes.

Based on our preliminary work we have shown how useful information for decision making could be extracted from a data warehouse and how it in turn could help to formulate the correct policies.

In this process we have highlighted the need to maintain graduate information. Further our selected examples were based on district, year and course as the dimensions. It is important to note that the star schema of figure 4 has four more dimensions as well as it could be further expanded by integrating data such as unemployment. Thus through creating the appropriate cubes, we can extract knowledge necessary to achieve our objectives.

We expect to further research this area before forwarding our recommendations. We need to investigate the existence of hidden data patterns and for that we need to apply data mining techniques and advanced queries.

6.0 ACKNOWLEDGMENTS

We acknowledge the support given by the departments of Education Examination and Census, UGC, Faculty of Medicine, Universities and UCSC.

7.0 REFERENCES

1) Chaudhuri S., Dayal U. "An Overview of Data Warehousing and OLAP Technology", SIGMOD Record 26:1, Mar 1997, pp 65-74.

2) Chawatte S.S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J., and Wisdom J., "The TSIMMIS Project: Integration of Heterogeneous Information Sources", Proc. of IPSJ Conf., Tokyo, Japan, Oct. 1994, pp. 7-18.

3) Han J. & Kamber M., "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

4) Harinarayan V., Rajaraman A., Ulman J.D. "Implementing Data Cubes Efficiently", Proc of SIGMOD Record, 25:2, 1996, pp. 205-216.

5) MOTET, Report of analysis of the graduate registration forms, Ministry of Tertiary Education & Training, Sept. 2002.

6) UGC, Report of the committee appointed to review university admissions policy, May 1984.

7) UGC, Report of the committee appointed to review university admissions policy, Dec. 1987.

8) UGC, Admission Booklet – Academic Year 2003-2004, 2003.

9) UGC undergraduate statistics. http://www.ugc.ac.lk

10) Wisdom J., "Research Problems in Data Warehousing", Proc. 4th Intl., CIKM Conf., 1995.