DECISION SUPPORT FOR EVERYONE: OLAP IN MS EXCEL

Gaj Vidmar, Emil Hudomalj
Institute of Biomedical Informatics
University of Ljubljana, Faculty of Medicine
Vrazov trg 2, 1000 Ljubljana, Slovenia
Tel: +386 1 5437770; fax: +386 1 5437771
e-mail: {gaj.vidmar,emil.hudomalj}@mf.uni-lj.si

ABSTRACT

Over the last ten years, Online Analytical Processing (OLAP) has become a very popular tool for interactive analysis of multidimensional information. Providing online operation and flexible summarising, tabulating and charting options, it has become an essential part of the decision support process in corporate setting. Our aim is to demonstrate how easy and effective it is to implement basic OLAP without any expert knowledge or high-priced software, using just the most common spreadsheet program. On a standard personal computer with MS Windows platform, we implemented a demonstrational application with public health data in MS Excel 2000, without any programming. After data cleansing the flat-file worksheet is instantly converted into an OLAP application with the user-friendly pivot table technology. A bonus of this approach is that the results can be made directly accessible over the WWW by publishing the workbook to a web server. Provided that the user has MS Internet Explorer and MS Office 2000 installed, all the drill-in, drill-out and dimension-swapping capabilities are accessible within the browser, while the data source remains fully protected. If the application is properly designed, privacy constraints are respected since all the information is only provided at the aggregate level. Contrary to widespread belief, storage and processing capabilities are not a serious issue with this approach with up to tens of thousands of records.

1 INTRODUCTION

In the early eighties, new methodologies for searching information in existing databases were developed, including knowledge discovery in databases (KDD) and online analytical processing (OLAP). A popular definition of KDD is "*the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data*" (Fayyad et al., 1996), while any further review of KDD is far beyond the scope of this article.

The term OLAP was introduced almost a decade ago in a report commissioned by a software vendor (Codd et al., 1993), but a less controversial contemporary definition is the acronym FASMI (*Fast Analysis of Shared Multidimensional Information)*, endorsed by N. Pendse (2002), a leading expert in the field. In addition to multidimensionality of the data, key features of OLAP are online operation, built-in and programmable analytical capabilities, and different presentational and reporting options. Common characteristics of KDD and OLAP algorithms are that they operate on large datasets and that their result, which mainly consists of aggregated information from existing databases, is previously unknown. The key advantage of OLAP over relational database management systems (RDBMS) and ordinary tables is interactive browsing of multidimensional and hierarchical data, while OLAP can also aid data integrity checking and reporting.

First OLAP implementations were limited to large enterprise and scientific datasets handled by proprietary systems. Today, numerous commercial systems are available and almost all RDBMS and statistical packages include support for OLAP. This powerful technology is hence available to anyone dealing with large datasets, but it comes at considerable price – not only in terms of software, but also or even more so in terms of implementation/consultancy costs.

2 EXCEL'S KEY FEATURES FOR DATABASES AND OLAP

Since its introduction in 1987, Excel has developed into the most popular and versatile spreadsheet application on the software market. Regarding database capabilities, the major developments were the introduction of multi-sheet workbooks and pivot tables in 1993 (version 5), new VBA interface and data validation in 1995 (version 8/97) and pivot charts in 1999 (version 9/2000).

2.1 Excel and databases

The starting point for database functionality in Excel is that any worksheet or part of a worksheet can serve as database once a header row of field names is followed by data rows below. It is strongly advised to use entire worksheets as data tables in any real-life application, though, especially because this enables the extremely simple yet powerful use of the AutoFilter feature, while separate sheets should be used for reports, pivot tables and charts. With Advanced Filters, of course, selection and searching is extended and refined, while simple Conditional Formatting instantaneously adds visual analytical functionality to any table, regardless of the level of data aggregation.

Being oriented towards MS Office products, this article cannot avoid the issue of Excel vs. Access. In spite of obvious limitations (quantity of data, one user at the time), Excel has crucial advantages over Access regarding reporting. As soon as detailed formatting and/or complex analyses are required in a report, or the report data should be used to generate other reports, Excel is the only option. Generally speaking, the analogy that Excel vs. Access is like car vs. truck is an extremely informative summary of the comparison – in terms of developers ("drivers"), speed, capacity, costs, adaptability and required organisational support. And just to complete the analogy, one can think of Oracle, SQL Server or SAP as trains, boats, airplanes.

Since this article aims at stressing what the user can do with Excel without any programming, using only built-in capabilities and bundled add-ins, we should mention three worksheet functions, accessible via simple formulas, which provide a true break-through in terms of database functionality: SUMPRODUCT, INDEX and MATCH. They are primarily meant for reporting, but they can be essential in the pre-processing phase of an OLAP application, or they can be used to provide the data to be subsequently processed by a pivot table.

Another database-related tool, which is a part of the standard MS Office package, is MS Query. Because it is not a separate application in the latest versions, it is even more hidden from the typical user; however, like the other components of the package, it has become a truly professional and powerful tool over the years. We have no space to even briefly address it in this paper, but we have to stress that its relatively simple operation provides Excel workbooks with full functionality of a relational database.

2.2 Pivot Tables and Pivot Charts

Pivot Tables are the key feature for doing OLAP in Excel without VBA programming. In a stepwise wizard-guided process, the user defines the data source and the fields that define pages, columns, rows and data (i.e., measures) of the pivot table. Various table options can be set along the way or after the table had been created. Instead of giving any further description of the topic, we refer the reader to Excel's extensive help and the web resources listed below.

Pivot Charts are designed using the same wizard as the Pivot Tables. We do not present an example here, because no simple visualisation would be truly informative and more appropriate than a table for our example data. However, there are cases when Excel's pivot charts are a useful OLAP tool, provided that they are designed by a person with sufficient background in fundamentals of statistics and data presentation.

2.3 Web resources

The use of Excel for data warehousing, reporting and OLAP can considerably reduce IT costs also because of the minimum training costs for the developers. A huge body of instructions and tips is publicly accessible thanks to a number of enthusiasts who maintain extremely valuable resource websites dedicated to Excel. Here we list a selection, ordered alphabetically by author surname:

- J. F. Lacher
  http://lacher.com/toc/tutpiv.htm
- P. Leclerc
  http://www.excel-vba.com/
- T. Mehta
  http://www.tushar-mehta.com/
- C. Pearson
  http://www.cpearson.com/excel.htm
- J. Peltier
  http://www.geocities.com/jonpeltier/Excel/index.html
- D. Steppan
  http://geocities.com/dsteppan/ExcelTop.html
- J. Walkenbach
  http://www.j-walk.com/ss/excel/index.htm

3 EXAMPLE APPLICATION

We present an example application from the field of public health statistics. The data were compiled from different sources for the purpose of exploring the relations between causes of death (according to ICD-10) and socio-economic characteristics (educational level, marital status, profession, etc.) for selected years 1992, 1995 and 1998 in Slovenia as part of our research collaboration with Barbara Artnik, MD, MSc, from the Institute of Social Medicine of the Faculty of Medicine in Ljubljana.

The initial data-cleansing phase consisted of elimination of incorrect entries, duplicates and inconsistencies based on exploratory statistical methods.

We designed two separate pivot tables, each with a single measure, in order to avoid formatting problems Excel exhibits when the multiple measures feature is used. In Figure 1, the use of several row and column dimensions is demonstrated and the categorical outcome is reported as

percentage using the standard trick of counting the number of cells with non-empty ID field. The active window demonstrates how filtering works.

In Figure 2, the measure is the average of a numerical field. Instead of containing zero value, empty cells are clearly marked, which is a straightforward option. The selected cell is aimed at demonstrating how drill-in and drill-out is achieved by double-clicking. The usefulness of the "preserve formatting" option is evident from both figures.

.

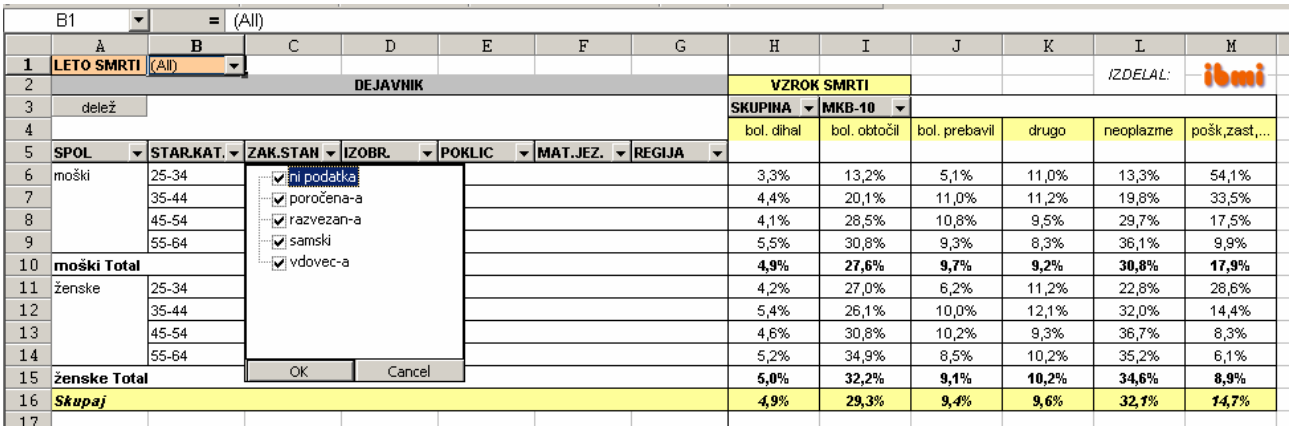

Figure 1: *Sample pivot table screenshot– measure is percentage of cases (public health statistics – mortality data).*

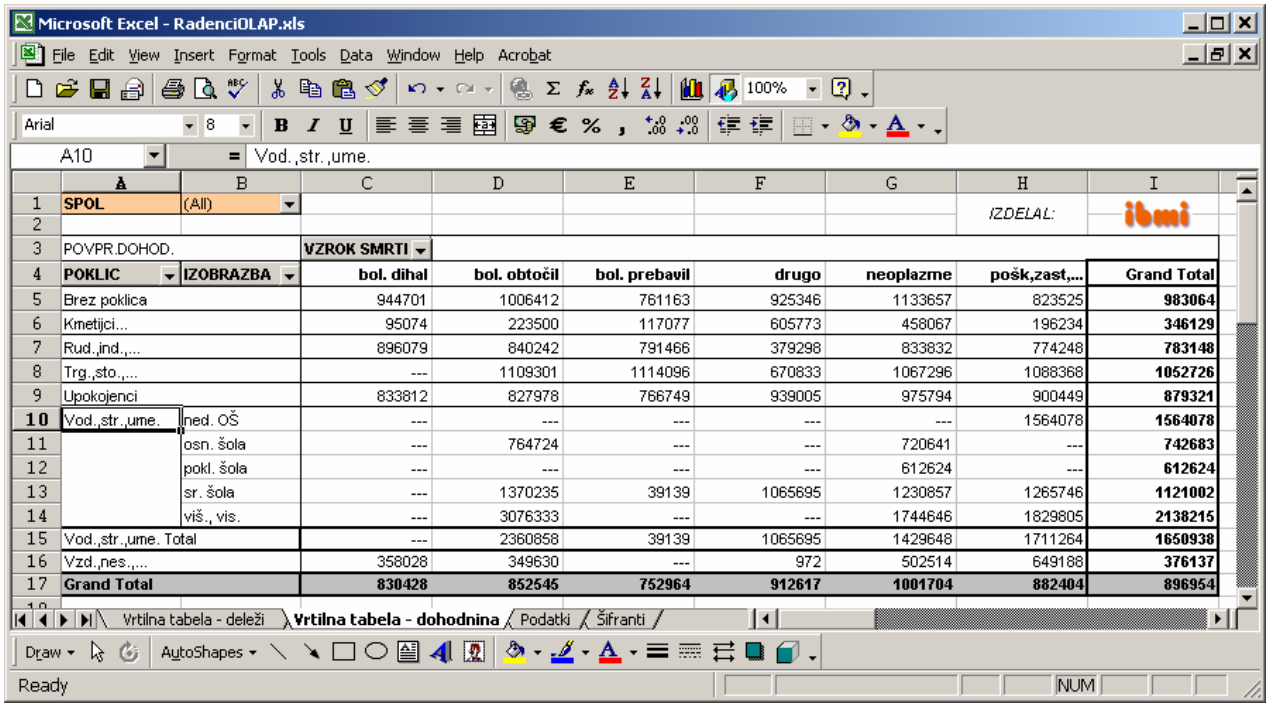

Figure 2: *Sample pivot table screenshot – measure is average of a numeric field (public health statistics – mortality data).*

Even though MS Excel 2000 is included or at least mentioned in his review, Thomsen (2002) argues extensively that spreadsheets cannot provide adequate OLAP functionally. He claims that they completely fail to meet four core OLAP requirements: multiple dimensions,

hierarchies, dimensional calculations and separation of structure and representation.

Rather then getting into a lengthy argumentation, we encourage anyone interested in challenging this opinion to try out Excel's capabilities and see for himself/herself that this is not really the case. In our opinion, the only serious problem are hierarchies. Hence, it is sensible to conclude that simple OLAP applications with a limited number of dimensions can be adequately, quickly and easily implemented in Excel. Needless to say, any such application is only valid to the extent to which the general issues with data for decision support (e.g., Poe, 1995) are properly dealt with.

A bonus of this approach is that the results can be made directly accessible over the WWW by publishing the workbook to a web server. Provided that the user has MS Internet Explorer and MS Office 2000 installed, all the drill-in, drill-out and dimension-swapping capabilities are accessible within the browser, while the data source remains fully protected. If the application is properly designed (data sheets hidden), privacy constraints are respected since all the information is only provided at the aggregate level. Contrary to widespread belief, storage and processing capabilities are not a serious issue with this approach with up to tens of thousands of records.

4 CONCLUSION

At virtually no cost, a functional OLAP application can be developed with MS Excel, based on the Pivot Table/Pivot Chart facility.

Database and OLAP functionality in Excel is a widely accessible technology and we believe that do-it-yourself decision-support systems based on it could find application in accounting, actuarial work, retail business, official statistics and elsewhere.

References

Codd E.F., Codd S.B., Salley C.T. (1993). Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. [Codd & Associates Technical Report, for Arbour Software, now Hyperion Solutions White Paper]

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). *From Data Mining to Knowledge Discovery and Data Mining: An Overview*. In Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (Eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1-34.

Pendse, N. (2001). *What is Olap?* [URL http://www.olapreport.com/fasmi.htm, part of The OLAP Report website]

Poe, V. (1995). *Building a Data Warehouse for Decision Support*. Upper Saddle River, NJ: Prentice Hall.

Thomsen, E. (2002). *OLAP Solutions: Building Multidimensional Information Systems* (Second Edition). New York: John Wiley.