# Analyzing Polls and News Headlines Using Business Intelligence Techniques

*Eleni Fanara, Gerasimos Marketos, Nikos Pelekis and Yannis Theodoridis*
*Department of Informatics, University of Piraeus,*
*80 Karaoli-Dimitriou St., GR-18534 Piraeus, Greece*
*efanara@gmail.com, {marketos, npelekis, ytheod}@unipi.gr*

## Abstract

Opinion and market research companies gather a substantial amount of polls data which can be combined with news headlines, for the corresponding time periods they are collected. These data are analyzed in order to answer specific (predefined) questions related to the situation of each time period. However, when these tasks are fulfilled, the collected data are archived and possibly the majority of them will remain unutilized for future research. In this paper, we argue that these 'inactive' data can be further analyzed and hidden knowledge can be extracted. For this reason, we propose an appropriate framework based on modern *Business Intelligence* (BI) techniques.

The innovation of the proposed framework is that it is able to reuse and analyze data that have been collected in the past and discover hidden knowledge, which can be utilized to bring profit in many ways. The basic scope of our framework is a) to supply knowledge on trends regarding specific politico-social and market issues and the way in which the situation of each period affect those trends, b) to predict the evolvement of trends according to the current situation or the formulation of emerging trends.

**Keywords: Business Intelligence, Knowledge Management, Opinion Research, Trend Analysis**

## 1. Introduction

A critical factor for the success of the modern enterprise is its ability to take advantage of the constantly increasing volume of information, coming from both internal and external sources (Cody et al, 2002). The successful enterprise exploits methodologies and tools for analyzing information and transforming it to knowledge in order to better support decision making processes.

Two technologies that have been proposed for managing information glut and offering a real competitive edge are Business Intelligence and Knowledge Management (KM). Business Intelligence developed a few years ago as a set of applications and technologies for gathering, storing, analyzing, and providing access to data to give a quality input to the decision making process. Knowledge Management is a broad term but in this paper we use it to describe technologies used for the management and analysis of unstructured information.

Market and opinion research industry is an interesting example to study the application of the above technologies in managing the information flooding phenomenon. Until today, such companies collect and store data both in structured (polls) and unstructured (news headlines) form but they make only limited use of them and only for the corresponding time periods they are collected.

However, latest technological advances have modified the definition of market research to one far more driven by technology, automation and speed than ever before. The new challenge in market research firms is to help business to thoroughly understand the implications of the trends that shape the global landscape and to develop competitive offerings, both in terms of products and the way they service their client's business (ESOMAR, 2006).

Therefore, the target is clear: research companies need a systematic approach to collecting, managing and analyzing information and transform it to knowledge on trends regarding specific politico-social and market issues and the way in which the situation of each period affect those trends. In this paper we describe how all these can be achieved using modern data management technologies.

The rest of the paper is organized as follows. In Section 2, we present core analytical techniques. Sections 3 and 4 present a framework for discovering knowledge on the market and opinion research industry and an application in a case study, respectively. Conclusions and hints for further work are drawn in Section 5.

## 2. Business Intelligence Technology

Business Intelligence developed a few years ago as a set of applications and technologies for gathering, storing, analyzing, and providing access to corporate data to aid in decision making. BI includes, among others, decision support systems (DSS), statistical analysis, information visualization, data warehousing (DW) and online analytical processing (OLAP), and data mining (DM).

In the following paragraph we outline the last three techniques as core analytical tools that will be utilized in the proposed framework.

### 2.1 Data Warehousing and OLAP analysis

Data warehousing is defined as a process of centralized data management and retrieval. It is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Major technological advances are making this vision a reality for many companies. And, equally major advances in data analysis software are allowing users to access this data freely.

Traditional Database Management Systems (DBMS) are known as operational database or OLTP (on-line transaction processing) systems as they support the daily storage and retrieval needs of an information system. Apart from querying, they support three main operations (insertions, updates and deletions) that can be formalized and executed over a DBMS using a Structured Query Language (SQL).

Nevertheless, maintaining summary data in a data warehouse can be used for data analysis purposes. Two popular techniques for analyzing data and interpreting their meaning are OLAP analysis and data mining. An important aspect in decision making is the level of details that the decision maker needs. Middle and upper management make complex and important decisions and therefore detailed data can not satisfy these requirements. Summarized data and hidden knowledge acquiring from the stored data, can lead to better decisions.

Additional to (naïve or advanced) database queries on detailed data, a data warehouse approach utilizes on-line analytical processing (OLAP). This is achieved by defining a cube following the multidimensional model (Agarwal et al, 1996). A data cube consists of a fact table containing keys to dimension tables and a number of appropriate measures. Dimension tables might have several attributes in order to build multiple hierarchies so as to support OLAP analysis whereas measures are used for analytical purposes (e.g., number of respondents, rate of optimism etc.). For each dimension we define a finest level of granularity which refers to the detail of the data stored in the fact table.

We illustrate the benefits obtained by such an approach with two examples of operations supported by a data warehouse build on polls and news headlines data and OLAP technologies:

- A user may ask to view the rate of optimism for the national economy among respondents with the same demographic profile e.g. age group, gender, marital status, from all over Greece (roll-up) or from specified areas even selecting if the area is urban or rural (drill-down).
- Given the ability to combine heterogeneous data a user may ask to view for a specific time period, news subjects with the greatest coverage that appeared amongst the first two headlines in the six TV stations (or in specified ones) and how that affected respondents answers on their opinion on the most important issue that the country is currently facing.

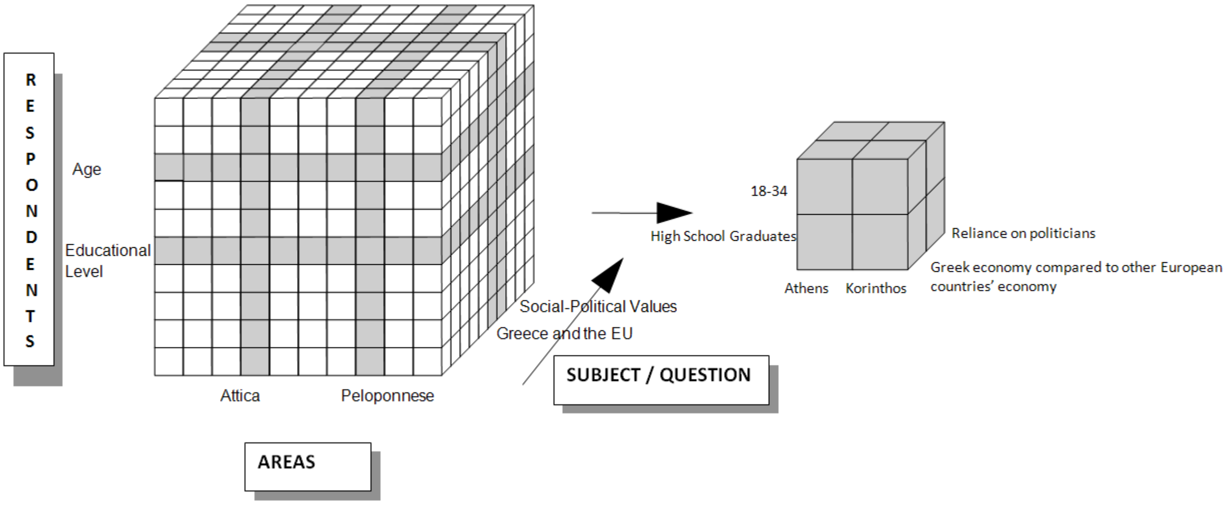


***Figure 1: Selecting parts of a cube by filtering a single (slice) or multiple dimensions (dice)***

Further to roll-up and drill-down operations described above, typical data cube operations include *slice* and *dice*, for selecting parts of a data cube by imposing conditions on a single or multiple cube dimensions, respectively (Figure 1), and *pivot*, which provides the user with alternative presentations of the cube (Figure 2).

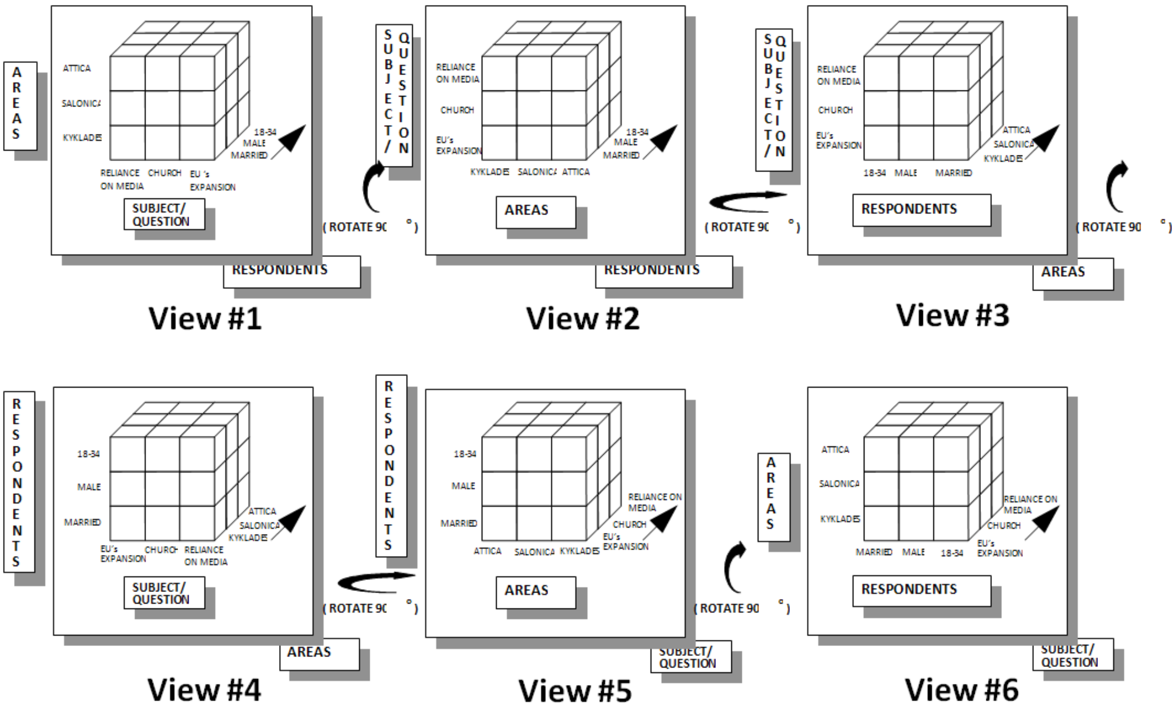


*Figure 2: Alternative presentations – views of a cube (pivot)*

## 2.2 Data mining

Applying mining techniques into a polls and news headlines data warehouse ultimately aims to the discovery of interesting, implicit and previously unknown knowledge. The *Knowledge Discovery in Databases* (KDD) process consists of the following steps, from the storage of interesting information in a data warehouse until the extraction, interpretation and understanding of useful, possibly hidden knowledge (Fayad et al., 1996), (Han & Kamber, 2000):

1. building a data warehouse from one or more raw databases (data warehouse building step);
2. selecting and cleansing data warehouse contents to focus on target data (selection and cleansing step);
3. transforming data to a format convenient for data mining (transformation step);
4. extracting rules and patterns by using data mining techniques (data mining step);
5. interpreting and evaluating data mining results to produce understandable and useful knowledge (interpretation and evaluation step)

Examples of useful patterns found through KDD process include clustering of respondents' profile (e.g. based on demographics - respondents whose family members are more than four and age is between 35 and 44 years old, based on political trends – respondents supporting governments' actions regarding economy matters), classification of trends regarding specific political or social matters, detecting trends semantics by using pattern finding techniques (e.g. measuring the impact of various news subjects on respondents answers regarding specific matters, based on the amount of media coverage they get, etc.). Recently, there have been proposals that expand the application of knowledge discovery methods on multi-dimensional data (Koperski & Han, 1995), (Koperski et al, 1998).

There are a lot of data mining algorithms that have been proposed in the literature. In the following paragraphs, we present the three most important categories:

Association rule mining aims at discovering interesting correlations among database attributes (Agrawal et al., 1993). Association rules are implications of the form A → B [s, c], $A \subset J$, $B \subset J$ where A, B and J are sets of items (i.e. attributes), characterized by two measures: support (s) and confidence (c). The support of a rule A → B expresses the probability that a database event contains both A and B, whereas the confidence of the rule expresses the conditional probability that a database event containing A also contains B. An example of the application of this kind of algorithm would be the detection of relation between apparently not related attributes like the respondents family members number and their opinion regarding the state of the country's economy compared to its last year's state.

Data clustering (Kaufman & Rousseeuw, 1990), (Jain et al., 1999) is the unsupervised process of grouping together sets of objects into classes with respect to a similarity measure. Thus, it is the behavior of groups rather than that of individual events that is detected. As an example, an analyst may ask for different clusters of respondents based on demographics that state that the political situation of the country is better or worse than last year's political situation, in periods that the media news present many news items regarding corruption of politicians.

Classification is one of the most common supervised learning techniques. The objective of classification is to first analyze a (labeled) training set and, through this procedure, build a model for labeling new data entries (Han & Kamber, 2000). In particular, at the first step a classification model is built using a training data set consisting of database records that are known to belong in a certain class and a proper supervised learning method, e.g. decision trees or neural networks. In case of decision trees, for example, the model consists of a tree of "if" statements leading to a label denoting the class the record it belongs in. At the second step, the built model is used for the classification of records not included in the training set. Many methods have been developed for classification, including decision tree induction, neural networks and Bayesian networks (Fayad et al., 1996). In our case, a user may ask to predict the Income level of the respondents that belong to a specific age group, are not married, and state that they are pessimistic regarding the economy of the country the next year.

## 3. A Framework for Discovering Knowledge

Essentially, our framework is able to reuse and analyze data that have been collected in the past and discover hidden knowledge, which can be utilized to bring profit in many ways. The basic scope of our framework is:

- To supply knowledge on trends regarding specific politico-social issues and the way in which the situation of each period affect those trends
- To predict the evolvement of trends according to the current situation or the formulation of emerging trends

*Figure 3* presents the proposed architecture that serves the task of collection of raw polls data and combines them with news headlines. Multiple excel files storing heterogeneous data are being processed using *Extract-Transform-Load* (ETL) packages in order to load transformed and cleansed data in the database. Summarized data will be stored in the data warehouse and end users will be able to query and analyze data using Business Intelligence techniques such as Data Mining and OLAP.

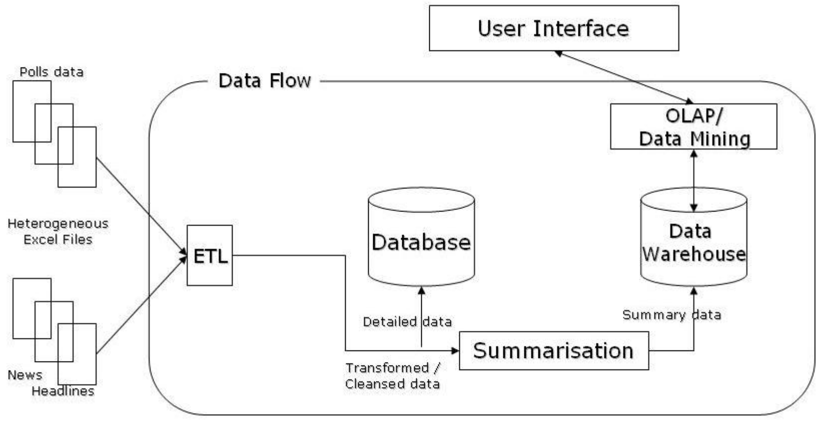


*Figure 3: The proposed architecture for analyzing polls and news headlines*

**Source data:** Polls and news headlines data are forming the input of the proposed architecture. Polls data files store information regarding all of the available variables (variables are considered the distinct questions of each survey). The news headlines data files store information regarding the headline, the broadcast station and the broadcast date. Apart from the raw polls data and the news headlines, a series of complementary data are required to build the database on which the architecture is based. The time periods, the TV stations, the polls data code frame (frame of codes corresponding to specific answers that respondents give), the news headlines categories are some of the required elements.
**Data Transformation and Loading:** Any necessary transformations take place in this stage. ETL packages undertake the tasks of the extraction and loading of the data to the database. An important part of this process is data cleansing, in which variations on schemas and data values from disparate data stores are resolved.
**Data Warehouse:** Summarized data construct the data warehouse which provides the integration of polls and news headlines. These data are in the appropriate form for further analysis.
**OLAP/Data Mining:** Based on the data warehouse OLAP and data mining techniques will be able to provide the users with the tools for discovering interesting patterns on the data. Respondents' segmentation according to various demographic data as well as their opinion on specific areas of interest and the impact of media on peoples' trends are typical examples of patterns likely to come up.

The proposed framework aims to discover hidden knowledge based on data that have already been processed and analyzed for various purposes and are no longer utilized. The use of already existing resources to produce new possible sources of profit is one of the key performance indicators for companies.

The innovation of the proposed framework lies on the alteration of the purpose of use of Business Intelligence tools which are not used for decision supporting procedures as they traditionally are, but for further utilizing existing resources of a company aiming to create new services for clients or even to give analysts one more scientific and innovatory way of documenting their findings and even predicting new trends based on the past data. In both cases the company is able to make profit and acquires a competitive advantage.

## 4. Applying Our Framework

We test the applicability and efficiency of our framework in a real world case study. We employ a set of six-year period (2001-2007) statistical data collected every four months during a pan-Hellenic political and social survey. We combine these data with news headlines that appeared in six TV stations during the survey periods. The polls data are provided by Metron Analysis, one of the top Greek research companies and the news data are provided by the Media Monitoring department of the same company.

### 4.1 Polls Data

The survey, conducted to gather our case study data, is based on a structured questionnaire which contains sections concentrating on various subjects. The sections usually contain both standard and ad-hoc questions based on the current situation. The completion of the questionnaire is based on specified routing of the questions asked to the respondent according to their previous answers. This means that not all of the respondents give answers to all of the questions. For example, if someone states that they are unfamiliar with a specific issue they will not be further asked on their opinion on the outcome of this issue.

The collected answers are of two types, open-ended and pre-coded. Open-ended answers are the answers in which the respondent is able to answer whatever they want, whereas pre-coded answers are the ones that the respondent has the restriction to choose their answer from a specified list of answers. In both cases the respondents' answers are being encoded before the data are processed for analysis. The list of codes generated is the code frame mentioned before.

### 4.2 News Headlines Data

News headlines data originate from the main news broadcasts of six Greek TV stations for the corresponding to the survey conduction time periods. Their collection is not complicated; the news headlines are stored in separate files according to the station they were broadcasted from and the date of the broadcast. The news headlines are categorized according to the main issue argued about, in categories like: unemployment, economy, inflation, international economy matters, education, insurance, health, science, foreign affairs, drugs, environment etc. Each station's ranking of the news headlines is also available.

### 4.3 The Database Schema

*Figure 4* presents the proposed E-R diagram for our framework. The source data (polls and news headlines) are organized in many tables which with the appropriate relationships give a unified dataset, which can be used as the basis for the development of the data warehouse.

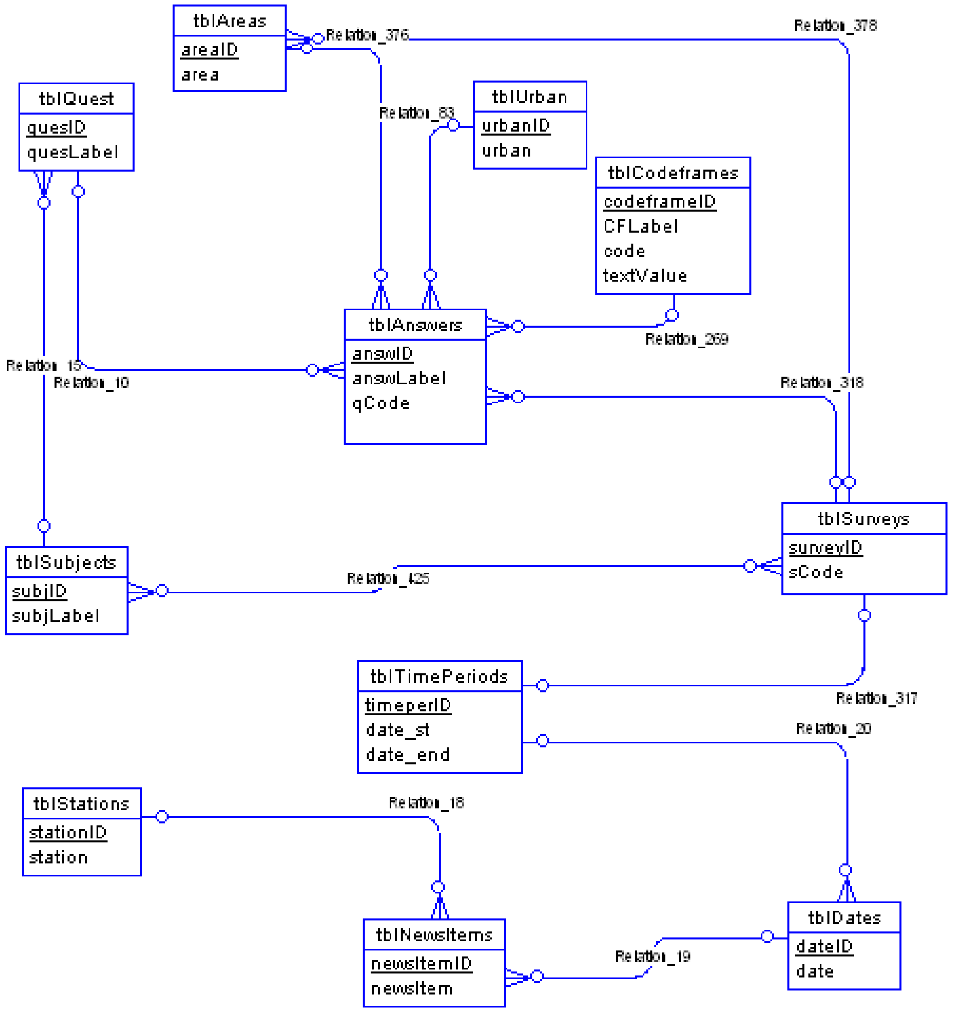


***Figure 4**: The E-R diagram of the database*

In short, the database includes tables *Answers*, *Questions*, *Subjects*, *Areas*, *Urban*, *Codeframes*, *Surveys*, *TimePeriods*, populated from polls data, tables *NewsItems* and *Stations* populated from news headlines data and a table of *Dates* to connect the two data sources.

## 4.4. The Data Warehouse

As mentioned before data warehouses are based on multi-dimensional data models. *Figure 5* shows the proposed architecture of our frameworks' data cube, which allows data to be modeled and viewed in multiple dimensions. The data cube is implemented using the snowflake's schema model, with two fact tables and a set of dimensional tables related with them.

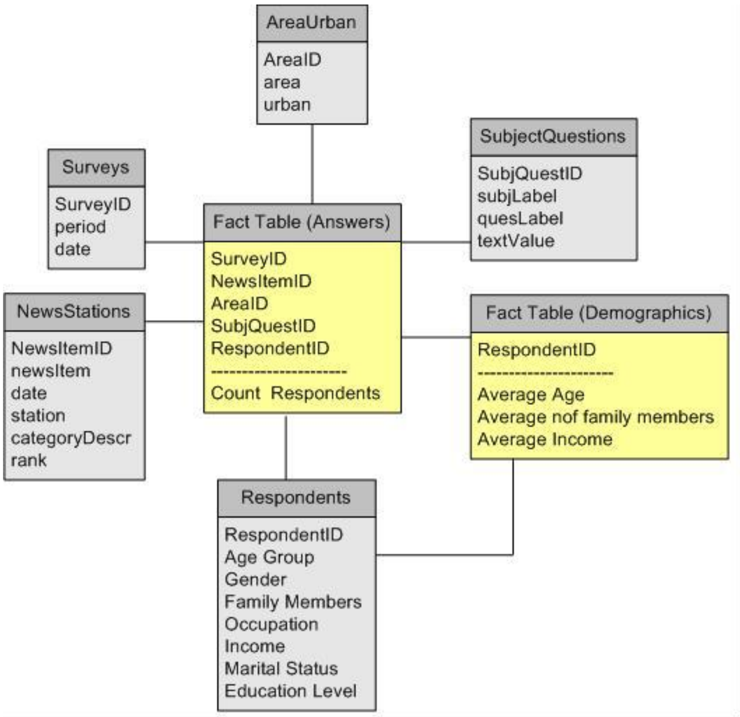
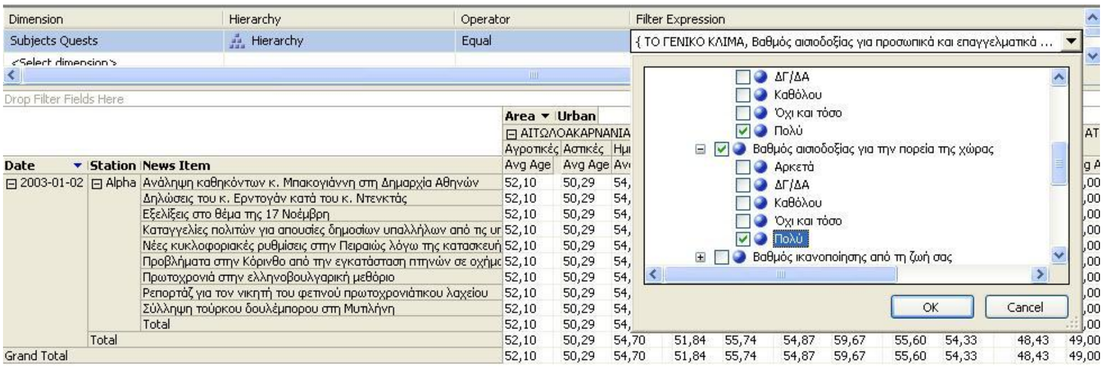


*Figure 5: A sample data warehouse for polls and news headlines data*

In order to extract useful knowledge the dimensions should maintain hierarchies like respondents' age group, educational level (and other demographics hierarchies), subjects and questions of the survey, news headlines and news categories etc, while the fact tables contain measures such as number of respondents, specific answer to specific question number, percentage of respondents answering to specific questions, percentage of news headlines regarding specific news categories, percentage of news headlines presented by specific TV stations.

In particular, dimension *AreaUrban* consists of a hierarchy representing urban and rural areas. Dimension *NewsStations* consists of a hierarchy for news headlines categorizing in specified news categories and a hierarchy for ranking the news headlines according to the order they were presented by each station. Dimension *Surveys* consists of a hierarchy that represents time periods and dates in which the survey was conducted. Dimension *Respondents* consists of many hierarchies each of which represents a different demographic variable. For example, there are hierarchies for Age, Gender, Educational Level, Income level, Insurance, Marital status, Nationality and Occupation. Dimension *SubjectsQuests* consists of a hierarchy which categorizes the questions of the survey to specific subjects.



| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Subjects Quests | Hierarchy | Equal | { ΤΟ ΓΕΝΙΚΟ ΚΛΙΜΑ, Βαθμός αισιοδοξίας για προσωπικά και επαγγελματικά ... |
| <Select dimension> | | | |

Drop Filter Fields Here

- ☐ ΔΓ/ΔΑ
- ☐ Καθόλου
- ☐ Όχι και τόσο
- ☑ Πολύ
- ⊟ ☑ Βαθμός αισιοδοξίας για την πορεία της χώρας
  - ☐ Αρκετά
  - ☐ ΔΓ/ΔΑ
  - ☐ Καθόλου
  - ☐ Όχι και τόσο
  - ☑ Πολύ
- ⊞ ☐ Βαθμός ικανοποίησης από τη ζωή σας

| | | | Area ▾ Urban | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ⊟ ΑΙΤΩΛΟΑΚΑΡΝΑΝΙΑ | | | | | | | | | | AT |
| | | | Αγροτικές | Αστικές | Ημι | | | | | | | | |
| Date | Station | News Item | Avg Age | Avg Age | Av | | | | | | | | g A |
| ⊟ 2003-01-02 | ⊟ Alpha | Ανάληψη καθηκόντων κ. Μπακογιάννη στη Δημαρχία Αθηνών | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Δηλώσεις του κ. Ερντογάν κατά του κ. Ντενκτάς | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Εξελίξεις στο θέμα της 17 Νοέμβρη | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Καταγγελίες πολιτών για απουσίες δημοσίων υπαλλήλων από τις υπ | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Νέες κυκλοφοριακές ρυθμίσεις στην Πειραιώς λόγω της κατασκευή | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Προβλήματα στην Κόρινθο από την εγκατάσταση πτηνών σε οχήμα | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Πρωτοχρονιά στην ελληνοβουλγαρική μεθόριο | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Ρεπορτάζ για τον νικητή του φετινού πρωτοχρονιάτικου λαχείου | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Σύλληψη τούρκου δουλέμπορου στη Μυτιλήνη | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | | Total | 52,10 | 50,29 | 54, | | | | | | | | ,00 |
| | Total | | 52,10 | 50,29 | 54,70 | 51,84 | 55,74 | 54,87 | 59,67 | 55,60 | 54,33 | 48,43 | 49,00 |
| Grand Total | | | 52,10 | 50,29 | 54,70 | 51,84 | 55,74 | 54,87 | 59,67 | 55,60 | 54,33 | 48,43 | 49,00 |

*Figure 6: A view of the processed data cube*

### 4.5. Data Mining

As we have already mentioned, association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used to explain respondents' trends towards various issues. *Table 1* illustrates the rules that came up describing some relationships between the size of families of respondents and other demographics and their opinion regarding the state of the country's economy compared to its last year's state

| # | Confidence | Rule |
|---|---|---|
| 1. | 0,59 | Age Group = 45-54, Gender = Women -> answer = Worse |
| 2. | 0,587 | Family Members < 3, Age Group = 45-54 -> answer = Worse |
| 3. | 0,587 | Family Members = 6 - 8, Gender = Women -> answer = Worse |
| 4. | 0,587 | Family Members = 6 - 8, Gender = Women -> answer = Worse |
| 5. | 0,581 | Age Group = 35-44, Gender = Women -> answer = Worse |
| 6. | 0,371 | Age Group = 18-34, Gender = Men -> answer = Same |
| 7. | 0,16 | Age Group = 65+, Gender = Men -> answer = Better |
| 8. | 0,15 | Family Members < 3, Gender = Men -> answer = Better |
| 9. | 0,15 | Family Members < 3, Gender = Men -> answer = Better |

***Table 1: Sample application of the Associations Rules algorithm***

Another even more interesting application of the Association Rules algorithm is the exploration of possible relation between news headlines categories at given periods and respondents answers regarding their optimism for their personal plans. By applying this mining model on the news and polls data warehouse we are able to confirm, an expected trend of respondents. Respondents with a lower educational level tend to answer negatively regarding their optimism for their personal plans when the news headlines greater coverage regards subjects like expensiveness, inflation, unemployment and difficulties in establishment in line of business. Table 2, illustrates the eight most interesting association rules in our sample data warehouse sorted on descending confidence and support.

| # | Confidence | Rule |
|---|---|---|
| 1. | 0,202 | NewsCateg = Inflation, Educational Level = College degree/More than College Degree, Gender = Men -> answer = Fairly optimistic |
| 2. | 0,199 | NewsCateg = Unemployment, Educational Level = College |

| | | |
|---|---|---|
| | | degree/More than College Degree -> answer = Fairly optimistic |
| 3. | 0,198 | NewsCateg = Unemployment, Age Group = 18-34, Educational Level = High School -> answer = Fairly optimistic |
| 4. | 0,124 | NewsCateg = Unemployment, Educational Level = Less than High School, Gender = Women -> answer = Not at all optimistic |
| 5. | 0,119 | NewsCateg = Expensiveness, Age Group = 65+, Educational Level = Less than High School -> answer = Not at all optimistic |
| 6. | 0,116 | NewsCateg = Inflation, Educational Level = Less than High School -> answer = Not at all optimistic |
| 7. | 0,115 | NewsCateg = Expensiveness, Age Group = 65+, Gender = Women -> answer = Not at all optimistic |
| 8. | 0,202 | NewsCateg = Inflation, Educational Level = College degree/More than College Degree, Gender = Men -> answer = Fairly optimistic |

***Table 2: Associating News Headlines with Polls data***

## 5. Conclusions

Taking into account the new definition of Market Research that technological advances have lately given to the industry, it is obvious that market research companies must stand at the peak of new technologies and should advance the services provided to their clients using all the offered means.

Data mining and OLAP techniques can provide a very reliable tool towards the above described aim. Using OLAP companies are able to integrate heterogeneous data coming from different data sources to a data warehouse and assisted by the data mining techniques they are able to discover ‘new’ knowledge or even confirm expected trends of people.

In this paper, we propose a framework for integrating news headlines data with opinion polls data. The main purpose of the framework is to show how a company using already existing resources and Business Intelligence technologies can achieve enhancement of the supplied services being at the same time innovative and cost effective.

## 6. Acknowledgements

We thank Metron Analysis for providing the polls and news headlines data on which our framework was tested. Gerasimos Marketos was supported by a PENED’2003 grant funded by the General Secretariat for Research and Technology of the Greek Ministry of Development.

## References

Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, R., & Sarawagi. S (1996), “On the computation of multidimensional aggregates”, *In Proceedings of 22th International Conference on Very Large Data Bases*, VLDB’96, *Bombay, India.*

Agrawal, R., Imielinski, T., & Swami, A. (1993), “Mining Association Rules between Sets of Items in Large Databases”, *In Proceedings of ACM SIGMOD International Conference on Management of Data, SIGMOD’93,* Washington DC, USA.

Cody, W.F., Kreulen, J. T., Krishna, V., & Spangler, W. S. (2002),”The integration of business intelligence and knowledge management”, *IBM Research Journal* 41(4)

Fayad, U., Piatetsky-Shapiro, G., Smith, P., & Uthurusami, R. (1996), *Advances in Knowledge Discovery and Data Mining*, MIT Press.

ESOMAR (2006), “Highlights 2006”, Technical Report.

Han, J., & Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

Jain, A., Murty, M., & Flynn, P. (1999), “Data Clustering: A Review”, *ACM Computing Surveys, 31(3).*

Kaufman, L., & Rousseeuw, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons.

Koperski K., J. & Han J. (1995), “Discovery of Spatial Association Rules in Geographic Information Databases”, *In Proceedings of the 4th International Symposium on Large in Spatial Databases, SSD’95*, Portland, MA, USA.

Koperski, K., Han, J., & Adhikary, J. (1998), “Mining Knowledge in Geographical Data”. *Communications of the ACM*, 26(1), 65–74.