Generation of interesting knowledge using data mining: a user-oriented approach[1]

Adolfo Guzman, Claudia Nogueron[2], Alicia Martinez[2]

[1] Center for Computing Research,
National Polytechnic Institute, Mexico
a.guzman@acm.org
[2] Technological Institute of Zacatepec, Morelos. Mexico
cnogo@hotmail.com, alimartin@dsic.upv.es

Abstract. Nowadays, the ongoing need to digest and understand enormous volumes of information is crucial for decision-making in economics, politics and science. One of the main issues associated to the large quantity of stored information is extract an interesting knowledge to specific purposes. One of the important problems that are used by data mining community is so-called classification of the information. In this paper, we propose four generic criteria to extract an interesting knowledge filter the information of a huge volume of databases. We use data mining techniques to accomplish the process to extract interesting knowledge based on user profiles.

Keywords. Interesting knowledge, Data mining, Remarkable winner-loser criterion, User type criterion, Generic time criterion, Pattern criterion.

1. Introduction

Nowadays, the use of information takes a radical relevance in different aspects of our society, such as politics, economics and science, where the immediate access to information is crucial in decision-making. As a consequence, this information dependency has brought different problems associated to the large quantity of stored information.

In this context data mining appears as a mean of *efficient discovering of valuable, non-obvious information from a great data collection* [1]. The main motivation of data mining is to offer people working with large amounts of data, the easiness to take the right decisions

[1] This work has been partially supported by the Asociación Nacional de Universidades e Instituciones de Educación Superior ANUIES and CONACYT, Mexico

based on those data. Data mining is based on the hypothesis that past data supplies a useful way to find valuable information. Data mining proposes mechanisms to analyze a large amount of data in order to present useful information to managers, directors, administrators and so on, in such a way they can make important decisions in companies, hospitals, schools, etc. The basic idea of data mining is to find hidden knowledge not noticed at first sight.

Some examples where data mining has been applied successfully are: a) telecommunication networks [2] b) sales on *Mexican Petroleum* (PEMEX) gas stations, and detection of anomalies in measurement instruments used in *Federal Commission of Electricity* (CFE) [1] d) extraction of expert decision-talking process from a clinical database [3] e) construction and maintenance of software components [4] f) non-supervised learning of relational patterns [5]. Despite the multiple advantages of the data mining techniques, and the great diversity of tools that make use of data mining [6] [7] there are some associated problems to their practical application. For example, current search systems rely on the users' ability to assemble the appropriate queries, which will analyze the database and retrieve useful knowledge. The main objective of this research work is address that problem by providing a methodology based on predefined criteria, in order to search for relevant knowledge in databases without user involvement. In this paper, we present that methodology emphasizing the criteria to search interesting and generic facts from a sample database.

The paper is structured as follows: Section 2 presents an experimental procedure, where the proposed criteria to search for interesting knowledge are detailed. Section 3 shows the results obtained with the application of the proposed criteria. Finally, Section 4 presents the conclusions.

2. Experimental Procedure

As previously stated, we have developed a methodology which supports the automatic discovery of interesting knowledge and abnormal trends, in a large set of database information. The four phases of our proposed methodology (Fig. 1) are summarized as follows:

Phase 1. Data extraction. This phase consists on creating a data cube starting from the relational database, selecting the axes of data according to our particular interest. An axis is a straight line of the cube that has a name according to the type of data it stores. The dimensions or coordinates of the cube refer to the number of axes; normally, a data cube contains three axes. The axes generally refer to products, geography and time, but they can also be of any other type. Once the information from the database has been extracted, it is placed in a tree

data structure. This is done for each axis of the data cube. In this proposal, the Bischoff methodology [8] is used for the creation of the data cube.

Phase 2. Exhaustive mapping. This phase consist on mapping all the information contained in the data cube. The mapping is done for each one of the axes in data cube, with all the other axes. If we have a data cube of three dimensions or axes we would have the following formula: Cube (A,B,C) = $A_1B_1C_1$, $A_1B_1C_2$, $A_1B_1C_3$, … , $A_1B_2C_1$, $A_1B_3C_1$, … , $A_KB_MC_{N..}$ Therefore the information produced by the exhaustive mapping will represent every possible combination of data contained in the data cube.

Phase 3. Classification. This phase consists on classifying the information obtained by the exhaustive generator. In this phase it is necessary to filter and classify the information contained in the data cube. In this research work, four criteria are proposed to filter information: 1) Remarkable winner-loser criterion, 2) User type criterion, 3) Generic time criterion, 4) Pattern criterion. In section 3, the proposed criteria are explained in detail.

Phase 4. Results visualization. This phase consists on storing interesting knowledge generated in the previous phase and graphically displaying them to users.

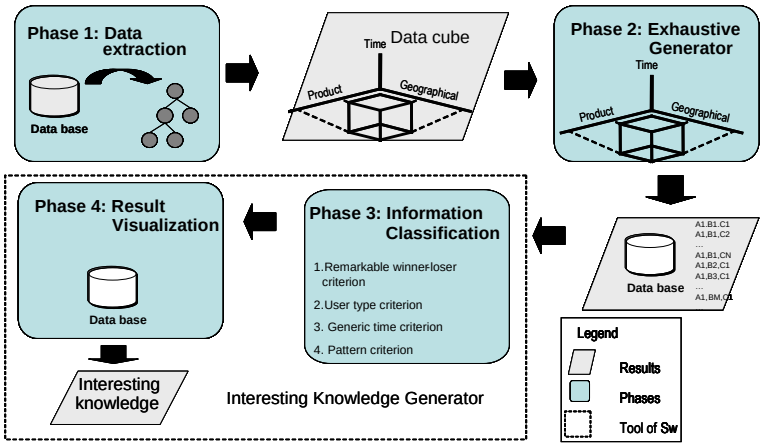


Fig. 1. The generator schema overview

2.1 Case study

In order to illustrate our approach, we analyze the database of *Foodmart,* a fictional and multinational chain of grocery stores, widely available from Microsoft SQL Server 2000 Analysis Services® installation. The information stored is related to products sold by supermarkets, sales dates, sales carried out in each branch, the quantity of sold products and the customers. The sample database has stored information of 1996 and 1997 and has 65 536 records. The data cube associated to this database contains 2 170 926 records and is constituted by the axes *products*, *geography* and *time*.

Table 1 shows a fragment of a cube obtained from database. The first column represents the country where the analyzed supermarket is located; the second and third column represent the state and city, respectively; the fourth column represents the year of the information record; the fifth column represents the month associated to the event; the sixth column represents the day associated to the event; the seventh column represents the brand; the eighth column represents the product sold on that branch on that date; the ninth column represents the quantity of the sold products on the analyzed date and branch.

Table 1. Data cube obtained from the database

Country	State	City	year	Month	day	Brand	product	Total of solds
Canada	BC	Victoria	1997	4	12	Top Measure	Top Measure Merlot Wine	4
Canada	BC	Victoria	1997	4	12	Tri-State	Tri-State Broccoli	18
Canada	BC	Victoria	1997	4	13	Tri-State	Tri-State Green Pepper	4
Canada	BC	Victoria	1997	4	14	Tri-State	Tri-State Lettuce	21
Canada	BC	Victoria	1997	4	14	Tri-State	Tri-State Macintosh Apples	3
Canada	BC	Victoria	1997	4	21	Best Choice	Best Choice Beef Jerky	2
Canada	BC	Victoria	1997	4	15	Tri-State	Tri-State Mixed Nuts	3
Canada	BC	Victoria	1997	4	15	Tri-State	Tri-State New Potatoes	5

2.2 The proposed criteria for looking for interesting knowledge

In this section we detail the proposed criteria for classifying the information obtained from the exhaustive mapping, from which every combination of items of data for columns in Table 1 has been generated.

1. Remarkable winner-loser criterion. This first criterion consists on filtering the information according to the numeric value of a column from the database, where this value stands out, positively or negatively, from the other records of the database. A record of the database will be considered winner if the value of the analyzed column stands out positively. We consider that a record is winner when the analyzed value is greater than the rest of the records in more than a 50%. In the opposite way, a record from the database is considered loser when the value of the analyzed column stands out negatively from the rest of the records of the database.

2. User type criterion. This criterion consists on filtering the information depending on the types of users that will have access to the database knowledge. It starts from the hypothesis that the information important for a user could not be important for another user. At this stage hierarchical levels are defined for showing the adequate information to each user. For example, in the supermarket's database, the general manager can be interested in visualizing

the information of the monthly sales reached by every department on the company. On the other hand, the manager of a certain department will be interested in visualizing the sold articles in his department, either daily or monthly.

3. Generic time criterion. This criterion consists on filtering the information considering the occurrence time of the stored incidences in the database, as well as the cyclical events of the database.

This criterion can be applied while searching for those events which occur on cyclical or seasonal intervals (Christmas, Easter, summer, holidays, etc.). Another information filter used in this criterion is the determination of statistical measures, like the average value of a record for a certain time interval. We have denominated this filter "*generic*". This way, we could have a *generic* month, or a *generic* weekday. The interesting knowledge generator uses these generic times to discover anomalies in the analyzed database. For example, this criterion allows us to analyze the store data in October. Specifically, this criterion allows us to compare the specific data of October with the generic October data, in other words, to compare all the information that happened during October for a certain number of years.

4. Pattern criterion. This criterion consists on filtering the information according to the abnormal behavior of an interesting knowledge. We use the concept of pattern to analyze the information that happens in a given time. Patterns are defined by Coplien: *"...the static and dynamic structures of solutions that occur repeatedly when producing applications in a particular context"* [9]. In this criterion, it is tried to visualize the generic behavior of the target record and discover the anomalous behavior it that could have. For example, if we analyze the generic behavior of a product, we would have the monthly sales of the same product; in this way it is possible to identify the similitude of the monthly product sales. An anomalous behavior is that where the sale of that product is different from the sales pattern of the other months. A fact is considered as an anomaly only when the difference of this behavior with the others overpasses a threshold of 10%. For example, let it be $vh_{2007}(t)$ the ice cream sales throughout 2007, we could find a 'generic curve' that consists on obtaining the minimum and maximum sales in January, and the same for each month. At the end we will have twelve intervals, one for each month. Let us think of an "acceptable channel or region" that happens in those twelve intervals. The ice cream sales in 2007, $vh_{2007}(t)$, will provoke interesting knowledge in the months where $vh_{2007}(t)$ goes out of the channel. A looser criterion will let us define as "interesting knowledge" only if $vh_{2007}(t)$ goes out of the channel in more than a 10%. This criterion is a great aid because it is applicable to all functions that show seasonality.

3. Results

In this section the main results, obtained from our approach after applying the proposed criteria in the case study, are shown.

1. *Remarkable winner-loser criterion.* The result obtained by applying this search criterion is shown in Fig. 2. The histogram is a partial view of database used (Table 1), which represents the amount of products sold. The analyzed record corresponds to the amount of sold products in a supermarket of Victoria, Canada on April 15th, 1997. The *'y'* axis indicates the analyzed product; the *'x'* axis indicates the amount of sold products in a branch. Therefore, the result generated for the case study, using this criterion, indicates that '*broccoli*' and '*lettuce*' are the most sold products in Victoria, Canada, and could be called '*winner products*'. On the other hand, the less sold product (loser) corresponds to '*apples*'.

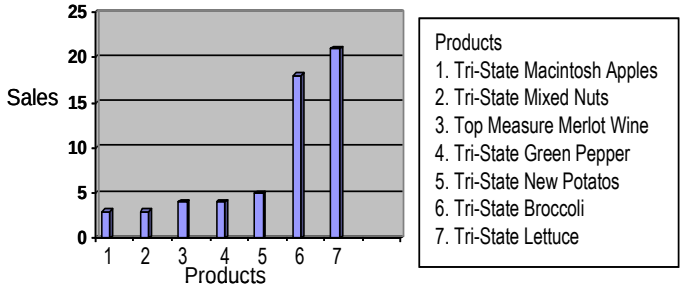


Fig. 2. Example of remarkable winner-loser criterion

2. *User type criterion.* The result obtained by applying this search criterion is shown in the Table 2. The generated view for the *General manager* contains information of all assigned departments to him (Top Measure, Tri-State, Best Choice) and the amount sold every April. In the case of the generated view for the manager of a certain department (Tri-State) shows the products sold in April of 1997, as well as the amount sold of every product.

Table 2. Example of user type criteria

General manager
(Top Measure,Victoria,April 1997,4)
(Tri-State,Victoria, April 1997,54)
(Best Choice,Victoria,April 1997,2)
Manager of department Tri-State
(Tri-State Broccoli,Victoria,12 April 1997,18)
(Tri-State Green Pepper,Victoria,13 April 1997,4)
(Tri-State Lettuce,Victoria,14 April 1997,21)
(Tri-State Macintosh Apples,Victoria,14 April 1997,3)
(Tri-State,Victoria,14 April 1997,24)
(Tri-State Mixed Nuts,Victoria,15 April 1997,3)
(Tri-State New Potatos,Victoria,15 April 1997,5)
(Tri-State,Victoria,15 April 1997,8)
(Tri-State,Victoria,April 1997,54)

3. *Generic time criterion.* The result obtained by applying this search criterion is shown in Table 3, which corresponds to the *generic* October month, the information generated in the sold articles in every October from 1995 to 1997. This is made according to a certain department and a specific place.

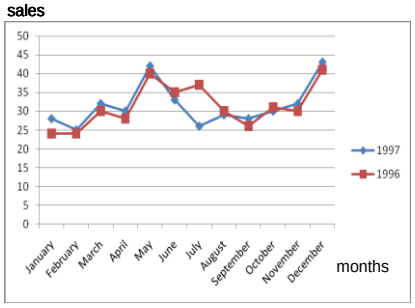
Table 3. Example of the generic time criterion

(Fast,Mexico,october1997,28)
(Fast,Mexico.october1996,23)
(Fast,Mexico,october 1995,21)
(Fast,Mexico. generic october,24)

4. *Pattern criterion.* The result obtained by applying this search criterion is shown in Fig. 3.a. The result shows the sales made of the Best Choice products in each month of 2007, where the values of February and December represent anomaly values that will overpass the threshold of 10% of the monthly sale pattern. The histogram shown on Fig. 3.b presents the sales made of the Best Choice in 1997 and 1996.

(Best Choice, Vancouver, January 1997, 28)
(Best Choice, Vancouver, February 1997, 25)
(Best Choice, Vancouver, March 1997, 32)
(Best Choice, Vancouver, April 1997, 30)
(Best Choice, Vancouver, May 1997, 42)
(Best Choice, Vancouver, June 1997, 33)
(Best Choice, Vancouver, July 1997, 26)
(Best Choice, Vancouver, August 1997, 29)
(Best Choice, Vancouver, September 1997, 28)
(Best Choice, Vancouver, October 1997, 30)
(Best Choice, Vancouver, November 1997, 32)
(Best Choice, Vancouver, December 1997, 28)

a) Results of the pattern criterion



b) The pattern criterion

Fig. 3. Example of sales in 1997 and 1996 for the brand Best Choice

4. Conclusions

In this paper, a methodology along with a set of generic criteria has been defined in order to find interesting knowledge from a database. The set of criteria are the basis for a data mining generator (which currently we are working on) based on a data cube. The proposed generator is a tool that allows the user, in an automatic way, to obtain interesting results from its database. The generator seeks interesting knowledge by applying the four proposed criteria. The search is done in each level of the corresponding tree to each axis of the cube, depending on if the obtained data are relevant for deeper searches in the tree.

The generator's notation is defined according to the cubes axes: (first axis, second axis, third axis, v), where 'v' could be 'large sale' or 'small sale'. In the case that the first axis is PRODUCTS, the second, GEOGRAPHY, and the third, TIME, then the notation would be as

follows: (PRODUCTS, GEOGRAPHY, TIME, v). The result shown by the generator through its notation, expresses the interests according to the user profile. Depending on the user profile, sometimes not all the criteria apply, but interesting results can be obtained with the applicable criteria. This means that if in the data cube the time axis does not exist, then the third criterion, generic time, could not be applied.

5. References

1. MARTINEZ L.G., GUZMAN A.A. Data Mining to search for patterns of behavior, International Symposium on Advanced Distributed Systems, Cucei,UDG. **2000** [in Spanish].
2. SASISEKHARAN R., SESHADRI V. Data Mining and forecasting in Large-Scale Telecommunications Networks. *IEEE Expert*. **11**, 1, pp. 37-43, **1996**.
3. TSUMOTO S. Extraction of Experts´ Decision Process from Clinical Databases Using Rough Set Model. *Proc. Firsth European Symposium: Principles of Data Mining and Knowledge Discovery,* pp. 58-67, **1997**.
4. BJORVAND A. Object Mining: A Practical Application of Data Mining for the Construction and Maintenance of Software Components. *Proc. Second European Symposyum: Principles of Data Mining and Knowledge Discovery,* pp. 121-129, **1998**.
5. SHEN W.M. LENG B. A Metapattern-Based Automated Discovery Loop for Integrated Data Mining Unsupervised Learning of Relational Patterns. *IEEE Trans. On Knowledge and Data Engineering*, **8**, 6, pp. 898-910, **1996**.
6. GEORGE H.J., YING Z. Mortgage Data Mining, in *Computational Intelligence in Financial Engineering. IEEE Press*, pp. 232-236. **1997**.
7. CLIFF B., JAMES K., RON K. MineSet: An Integrated System for Data Mining, **1997**.
8. BISCHOFF J. Data Warehouse, Design tables Multidimensional, Prentice Hall. p. 195, **1997**.
9. COPLIEN J. O., SCHMIDT D. C., Eds. Pattern languages of program design, ACM Press/Addison-Wesley Publishing Co., New York, NY, **1995**.