An Application of Case-Based Reasoning in Multidimensional Database Architecture*

Dragan Simić[1], Vladimir Kurbalija[2], Zoran Budimac[2]

[1] Novi Sad Fair, Hajduk Veljkova 11, 21000 Novi Sad, Yugoslavia
dsimic@nsfair.co.yu

[2] Department of Mathematics and Informatics, Fac. of Science, Univ. of Novi Sad
Trg D. Obradovića 4, 21000 Novi Sad, Yugoslavia
kurba@im.ns.ac.yu, zjb@im.ns.ac.yu

ABSTRACT. A concept of decision support system is considered in this paper. It provides data needed for fast, precise and good business decision making to all levels of management. The aim of the project is the development of a new online analytical processing oriented on case-based reasoning (CBR) where a previous experience for every new problem is taken into account. Methodological aspects have been tested in practice as a part of the management information system development project of "Novi Sad Fair". A case study of an application of CBR in prediction of future payments is discussed in the paper.

1 Introduction

In recent years, there has been an explosive growth in the use of database for decision support systems. This phenomenon is a result of the increased availability of new technologies to support efficient storage and retrieval of large volumes of data: *data warehouse* and *online analytical processing* (OLAP) products. A data warehouse can be defined as an online repository of historical enterprise data that is used to support decision-making. OLAP refers to technologies that allow users to efficiently retrieve data from the data warehouse.

In order to help an analyst focus on important data and make better decisions, case-based reasoning (CBR - an artificial intelligence technology) is introduced for making predictions based on previous cases. CBR will automatically generate an answer to the problem using stored experience, thus freeing the human expert of obligations to analyse numerical or graphical data.

The use of CBR in predicting the rhythm of issuing invoices and receiving actual payments based on the experience stored in the data warehouse is presented in this paper. Predictions obtained in this manner are important for future planning of a com-

* Research was partially supported by the Ministry of Science, Technologies and Development of Republic of Serbia, project no. 1844: "Development of (intelligent) techniques based on software agents for application in information retrieval and workflow"

pany such as the ”Novi Sad Fair” because achievement of sales plans, revenue and company liquidation are measures of success in business. Performed simulations show that predictions made by CBR differ only for 8% in respect to what actually happened. With inclusion of more historical data in the warehouse, the system gets better in predictions. Furthermore, the system uses not only a data warehouse but also previous cases and previous predictions in future predictions thus learning during the operating process.

The combination of CBR and data warehousing, i.e. making an OLAP intelligent by the use of CBR is a rarely used approach, if used at all. The system also uses a novel CBR technique to compare graphical representation of data which greatly simplifies the explanation of the prediction process to the end-user [3].

The rest of the paper is organized as follows. The following section elaborates more on motivations and reasons for inclusion of CBR in decision support system. This section also introduces our case-study on which we shall describe the usage of our system. Section three overviews the case based reasoning technique, while section four describes the original algorithm for searching the previous cases (curves) looking for the most similar one. Fifth section describes the actual application of our technique to the given problem. Section six presents the related work, while the seventh section concludes the paper.

2 User requirements for decision support system

“Novi Sad Fair” represents a complex organization considering the fact that it is engaged in a multitude of activities. The basic Fair activity is organizing fair exhibitions, although it has particular activities throughout the year. Ten times a year, 27 fair exhibitions are organized where nearly 4000 exhibitors take part, both from the country and abroad.

Besides designing a ‘classical’ decision support system based on a data warehouse and OLAP, requirements of the company management clearly showed that it will not be enough for good decision making. The decision to include artificial intelligence methods in general and CBR in particular into the whole system was driven by the results of the survey. The survey was made on the sample of 42 individuals (users of the current management information system) divided into three groups: strategic-tactical management (9 people), operational managers (15 people), and transactional users (18 people).

After a statistical evaluation of the survey [5], the following conclusions (among others) were drown:

- Development of the decision support systems should be focussed on problems closely related to financial estimates and financial marker trends tracking which span several years.
- The key influences on business (management) are political and economic environment of the country and region, which induces the necessity of exact implementation of those influences in the observed model (problem). Also it is necessary to take them into account in future events estimations.

- The behavior of the observed case does not depend on its pre-history but only on initial state, respectively.

Implementation of this non-exact mathematical model is a very complex problem. As an example, let us take a look into the problem pointed to us by company managers.

During any fair exhibition the total of actual income is only 30% to 50% of the total invoice value. Therefore, managers want to know how high the payment of some fair services would be in some future time, with respect to invoicing. If they could predict reliably enough what would happen in the future, they could make important business activities to ensure faster arrival of invoiced payments and plan future activities and exhibitions better.

The classical methods can not explain influences on business and management well enough. There are political and economic environments of the country and region that cannot be successfully explained and used with classical methods: war in Iraq, oil deficiency, political assassinations, terrorism, spiral growth in mobile telecommunication industry, general human occupation and motivation. And this is even more true in an enterprise such as Fair whose success depends on many external factors.

One possible approach in dealing with external influences is observing the case histories of similar problems (cases) for a longer period of time, and making estimations according to that observation. This approach, generally speaking, represents intelligent search which is applied to solving new problems by adapting solutions that worked for similar problems in the past - case-based reasoning.

3 Case based reasoning

Case-Based Reasoning is a relatively new and promising area of artificial intelligence and it is also considered a problem solving technology (or technique). This technology is used for solving problems in domains where experience plays an important role [2].

Generally speaking, case-based reasoning is applied to solving new problems by adapting solutions that worked for similar problems in the past. The main supposition here is that similar problems have similar solutions. The basic scenario for mainly all CBR applications looks as follows. In order to find a solution of an actual problem, one looks for a similar problem in an experience base, takes the solution from the past and uses it as a starting point to find a solution to the actual problem. In CBR systems experience is stored in a form of cases. The case is a recorded situation where problem was totally or partially solved, and it can be represented as an ordered pair *(problem, solution)*. The whole experience is stored *in case base*, which is a set of cases and each case represents some previous episode where the problem was successfully solved.

The main problem in CBR is to find a good similarity measure – the measure that can tell to what extent the two problems are similar. In the functional way similarity can be defined as a function $sim : U \times CB \rightarrow [0 , 1]$ where U refers to the universe of all objects (from a given domain), while CB refers to the case base (just those objects

which were examined in the past and saved in the case memory). The higher value of the similarity function means that these objects are more similar [1].

The case based reasoning system has not the only goal of providing solutions to problems but also of taking care of other tasks occurring when used in practice. The main phases of the case-based reasoning activities are described in the *CBR-cycle* (fig. 1) [1].

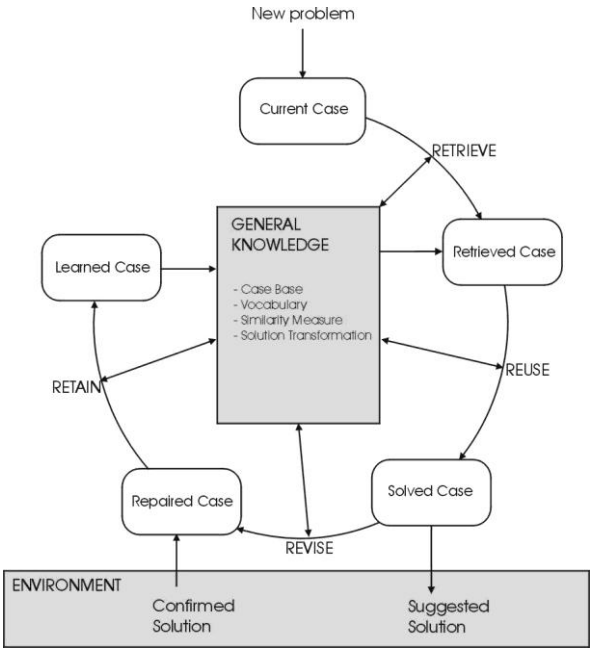

Fig. 1. The *CBR-Cycle* after Aamodt and Plaza (1994)

In the *retrieve* phase the most similar case (or *k* most similar cases) to the problem case is retrieved from the case memory, while in the *reuse* phase some modifications to the retrieved case are done in order to provide better solution to the problem (case adaptation). As the case-based reasoning only suggests solutions, there may be a need for a correctness proof or an external validation. That is the task of the phase *revise*. In the *retain* phase the knowledge, learned from this problem, is integrated in the system by modifying some knowledge containers.

The main advantage of this technology is that it can be applied to almost any domain. CBR system does not try to find rules between parameters of the problem; it just tries to find similar problems (from the past) and to use solutions of the similar problems as a solution of an actual problem. So, this approach is extremely suitable for less examined domains – for domains where rules and connections between parameters are not known. The second very important advantage is that CBR approach to learning and problem solving is very similar to human cognitive processes – people take into account and use past experiences to make future decisions.

4 CBR for predicting curves behaviour

The CBR system for its graphics in presenting both the problem and the cases is used [3]. The reasons are that in many practical domains some decisions depend on behaviour of time diagrams, charts and curves. The system therefore analyses curves, compares them to similar curves from the past and predicts the future behaviour of the current curve on the basis of the most similar curves from the past.

The main problem here, as almost in every CBR system, was to create a good similarity measure for curves, i.e. what is the function that can tell to what extent the two curves are similar. In many practical domains data are represented with the set of points, where the point is an ordered pair *(x,y)*. Very often the pairs are *(t,v)* where *t* represents time and *v* represents some value in the time *t*. When the data is given in this way (as a set of points) then it can be graphically represented. When the points are connected, then they represent some kind of a curve. If the points are connected only with straight lines then it represents the linear interpolation, but if someone wants smoother curves then some other kind of interpolation with polynomials must be used. There was a choice between a classical interpolating polynomial and a cubic spline. The cubic spline was chosen for two main reasons:

- Power: for the n+1 points classical interpolating polynomial has the power n, while cubic spline always has the power 4.
- Oscillation: if only one point is moved (which can be the result of bad experiment or measuring) classical interpolating polynomial significantly changes (oscillates), while cubic spline only changes locally (which is more appropriate for real world domains).

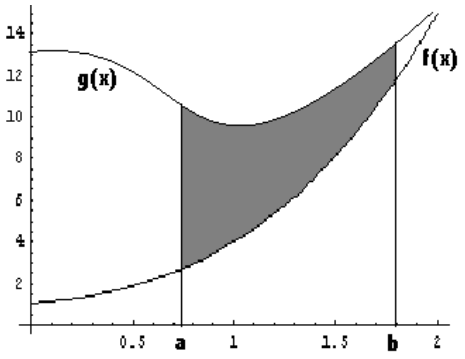

Fig. 2. Surface between two curves

When the cubic spline is calculated for curves then one very intuitive and simple similarity (or distance – which is the dual notion for similarity[1]) measure can be used. The

[1] When the dictance *d* i known then the similarity *sim* can be easily computed using for example function: *sim = 1/(1+d)*

distance between two curves can be represented as a surface between these curves as seen on the fig 2. This surface can be easily calculated using the definitive integral. Furthermore, the calculation of the definitive integral for polynomials is a very simple and efficient operation.

5 Application of the system

A data warehouse of "Novi Sad Fair" contains data about payment and invoicing processes from the last 3 years for every exhibition - containing between 25 and 30 exhibitions every year. Processes are presented as sets of points where every point is given with the time of the measuring (day from the beginning of the process) and the value of payment or invoicing on that day. It can be concluded that these processes can be represented as curves. Note that the case-base consists of cases of all exhibitions and that such a case-base is used in solving concrete problems for concrete exhibitions. The reason for this is that environmental and external factors influence business processes of the fair to a high extent.

The measurement of the payment and invoicing values was done every 4 days from the beginning of the invoice process in duration of 400 days, therefore every curve consists of approximately 100 points. By analysing these curves, the process of invoicing usually starts several months before the exhibition and that value of invoicing rapidly grows approximately to the time of the beginning of exhibition. After that time the value of invoicing remains approximately the same till the end of the process. That moment, when the value of invoicing reaches some constant value and stays the same to the end, is called the *time of saturation* for the invoicing process, and the corresponding value – the *value of saturation*.

The process of payment starts several days after the corresponding process of invoicing (process of payment and invoicing for the same exhibition). After that the value of payment grows, but not so rapidly as the value of invoicing. At the moment of exhibition the value of payment is between 30% and 50% of the value of invoicing. After that, the value of payment continues to grow to some moment when it reaches a constant value and stays approximately constant till the end of the process. That moment is called the *time of saturation* for the payment process, and the corresponding value – the *value of saturation.*

Payment time of saturation is usually couple of months after the invoice time of the saturation, and the payment value of saturation is always less than the invoice value of saturation or equal. The analysis shows that payment value of saturation is between 80% and 100% of the invoice value of saturation. The maximum represents a total of services invoiced and that amount is to be paid. The same stands for the invoicing curve where the maximum amount of payment represents the amount of payment by regular means. The rest will be paid later by court order, other special business agreements or, perhaps, will not be paid at all (debtor bankruptcy).

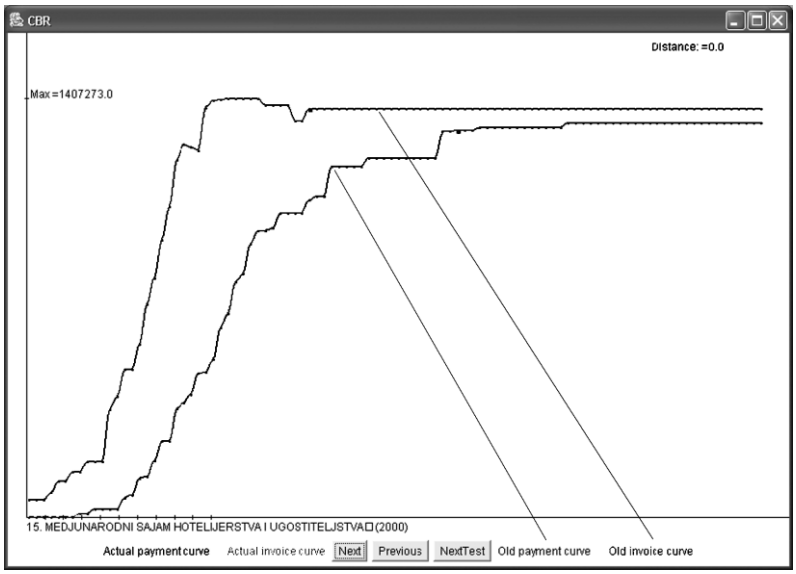

Fig. 3. The curves from the data mart, as the "Old payment curve" and the "Old invoice curve"

One characteristic invoice and a corresponding payment curve as the "Old payment curve" and "Old invoice curve" from the ”curve base” are shown (fig. 3). The points of saturation (time and value) are represented with the emphasised points on curves.

At the beginning system reads the input data from two data marts: one data mart contains the information about all invoice processes for every exhibition in the past 3 years, while the other data mart contains the information about the corresponding payment processes. After that, the system creates splines for every curve (invoice and payment) and internally stores the curves in the list of pairs containing the invoice curve and the corresponding payment curve.

In the same way system reads the problem curves from the third data mart. The problem is invoice and a corresponding problem curve at the moment of the exhibition. At that moment, the invoice curve reaches its saturation point, while the payment curve is still far away from its saturation point. These curves are shown as the "Actual payment curve" and the "Actual invoice curve" (fig. 4).

The solution of this problem would be the saturation point for the payment curve. This means that system helps experts by suggesting and predicting the level of future payments. At the end of the total invoicing for selected fair exposition, operational exposition manager can get a prediction from CBR system of a) the time period when payment of a debt will be made and b) the amount paid regularly.

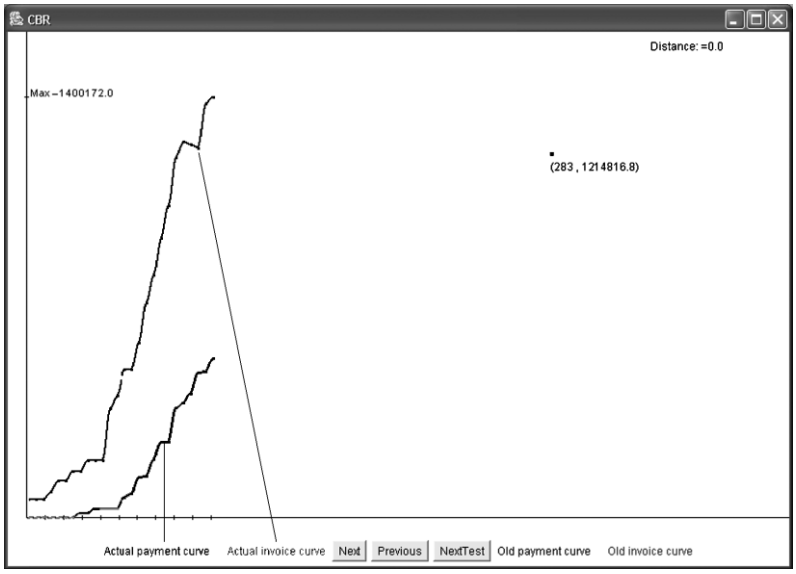

Fig. 4. Problem payment and invoice curves, as the "Actual payment curve" and the "Actual invoice curve" and prediction for the future payments

Time point and the amount of payment of a debt are marked on the graphic by a big red dot (fig. 4). When used with the subsets of already known values, CBR predicts the results that differed around 10% in time and 2% in value from actually happened.

5.1 Calculation of saturation points and system learning

The saturation point for one prediction is calculated by using 10% of the most similar payment curves from the database of previous payment processes. The similarity is calculated by using the previosly described algorithm. Since the values of saturation are different for each exhibition, every curve from the database must be scaled with a particular factor so that the invoice values of saturation of the old curve and actual curve are the same. That factor is easily calculated as:

$$Factor = \frac{actual\_value\_of\_saturation}{old\_value\_of\_saturation}$$

where the actual value of saturation is in fact the value of the invoice in the time of the exhibition.

The final solution is then calculated by using payment saturation points of the 10% most similar payment curves. Saturation points of the similar curves are multiplied with the appropriate type of goodness and then summed. The values of goodness are directly proportional to the similarity between old and actual curves, but the sum of all goodnesses must be 1. Since the system calculates the distance, the similarity is calculated as:

$$sim = \frac{1}{1 + dist}$$

The goodness for every old payment curve is calculated as:

$$goodness_i = \frac{sim_i}{\sum_{all\_j} sim_j}$$

At the end, the final solution – payment saturation point is calculated as:

$$sat\_point = \sum_{all\_i} goodness_i \cdot sat\_point_i$$

The system draws the solution point at the diagram combined with the saturation time and value. The system also supports solution revising and retaining (fig. 1). By memorizing a) the problem, b) suggested solution, c) the number of similar curves used for obtaining the suggestion and d) the real solution (obtained later), the system uses this information in the phase of reusing the solution for future problems. The system will then use not only 10% percent of the most similar curves but will also inspect the previous decisions in order to find 'better' number of similar curves that would lead to the better prediction.

6 Related work

The system presented in the paper represents a useful coexistence of a data warehouse and a case based reasoning resulting in a decision support system. The data warehouse (part of the described system) has been in operation in "Novi Sad Fair" since 2001 and is described in more details [5] [6] [7]. The part of the system that uses CBR in comparing curves has been done during the stay of the second author at Humboldt University in Berlin and is described in [3] in more detail.

Although CBR is successfully used in many areas (aircraft conflict resolution in air traffic control, optimizing rail transport, subway maintenance, optimal job search, support to help-desks, intelligent search on the internet) [4], it is not very often used in combination with data warehouse and in collaboration with classical OLAP, probably due to novelty of this technique. CBR does not require causal model or deep understanding of a domain and therefore it can be used in domains that are poorly defined, where information is incomplete, contradictory, or where it is difficult to get sufficient domain of knowledge. All this is typical for business processing.

Besides CBR, other possibilities are rule base knowledge or knowledge discovery in database where knowledge evaluation is based on rules [1]. The rules are usually generated by combining propositions. As the complexity of the knowledge base increases, maintaining becomes problematical because changing rules often implies a lot of reorganization in a rule base system. On the other side, it is easier to add or delete a case in a CBR system, which finally provides the advantages in terms of learning and explicability.

Applying CBR to curves and its usage in decision making is also a novel approach. According to the authors' findings, the usage of CBR, looking for similarities in curves and predicting future trends is by far superior to other currently used techniques.

7 Conclusion

The paper presented the decision support system that uses CBR as an OLAP to the data warehouse. The paper has in greater detail described the CBR part of the system giving a thorough explanation of one case study.

There are numerous advantages of this system. For instance, based on CBR predictions, operational managers can make important business activities, so they would: a) make payment delays shorter, b) make the total of payment amount bigger, c) secure payment guarantee on time, d) reduce the risk of payment cancellation and e) inform senior managers on time. By combining graphical representation of predicted values with most similar curves from the past, the system enables better and more focussed understanding of predictions with respect to real data from the past.

Senior managers can use these predictions to better plan possible investments and new exhibitions, based on the amount of funds and the time of their availability, as predicted by the CBR system.

Presented system is not only limited to this case-study but it can be applied to other business values as well (expenses, investments, profit) and it guarantees the same level of success.

Acknowledgement

The CBR system that uses graphical representation of problem and cases [3] was implemented by V. Kurbalija at Humboldt University, Berlin (AI Lab) under the leadership of Hans-Dieter Burkhard and sponsorship of DAAD (German academic exchange service). Authors of this paper are grateful to Prof. Burkhard and his team for their unselfish support without which none of this would be possible.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches, *AI Commutations*, pp. 39-58. 1994.
2. Zoran Budimac, Vladimir Kurbalija: Case-based Reasoning – A Short Overview, Conference of Informatics and IT, Bitola, 2001.
3. Vladinmir Kurbalija: On Similarity of Curves – project report, Humboldt University, AI Lab, Berlin, 2003.
4. Mario Lenz, Brigitte Bartsh-Sporl, Hans-Dieter Burkhard, Stefan Wess, G. Goos, J. Van Leeuwen, B. Bartsh: Case-Based Reasoning Technology: From Foundations to Aplications, Springer Verlag, October 1998.
5. Dragan Simic: Financial Prediction and Decision Support System Based on Artificial Intelligence Technology, Ph.D. thesis, draft text – manuscript, Novi Sad 2003.
6. Dragan Simic: Reengineering management information systems, contemporary information technologies perspective, Master thesis, Novi Sad 2001.
7. Dragan Simic: Data Warehouse and Strategic Management, Strategic management and decision support systems, Palic, 1999.