CHAPTER 25

Identifying Security Holes in OLAP Applications

JÜRGEN STEGER, HOLGER GÜNZEL, ANDREAS BAUER
Department of Database System
University of Erlangen-Nuremberg
Martensstr. 3
D-91058 Erlangen
Germany

Key words: Data Warehouse, OLAP, Multidimensional Data Model, Inference

Abstract: A data warehouse system is a necessity for fundamental decisions in every enterprise. The integration of data from several internal or external sources and the functionality of modern decision support systems like OLAP tools not only provide broad access to data but also raise security problems. Security concerns are more or less the same as those of other database systems but enriched especially with access and inference control in the multidimensional model. This paper presents an integrated approach for inference and access control not on the physical but on the conceptual level. The question is not only the restriction of relations, but rather the identification and evaluation of the inference problem of hierarchies and dimensions. The possibility to restrict or perturbate data in general, is not an adequate solution. We present some specific problems of a market research company and a solution with an indicator to discover possible attacks and so be able to restrict the access by mechanisms like aggregation, restriction or perturbation.

INTRODUCTION

Whenever managers of large enterprises prepare to make the right decisions, they require detailed information on specific topics. In a dynamic market environment it is crucial to have online information about one's own general business figures as well as detailed information on other companies. Some of the required data come from independent services like market research companies. They provide an additional basis for efficiently making decisions.

One of these third party data providers is the retail research department of GfK AG in Nuremberg. They collect the turnover of goods in different shops on a regular time base and sell this cleansed and aggregated data to companies in the form of trend analysis, market concentration and price class analyses. The market research companies themselves collect and store the provided data in a data warehouse [1]. Typically, users navigate through the data guided by a multidimensional data model which fits best for this application context. Codd introduced the term “Online Analytical Processing” (OLAP, [2]) for the interactive exploration of multidimensional data.

The goal of data warehousing and OLAP is to provide integrated access to data which resided in heterogeneous sources. Each user is able to analyze data and create reports himself. But on the other side the user is now able to receive data he is normally not allowed to. New access control mechanism are required, because the variety of data cause security problems. Access control depends on authorization of a specific user, i.e. not each user should be allowed to see the same data. Commercial products already restrict access to specific classification hierarchy nodes or provide only aggregated data or only parts of a cube.

Beside these static problems, topics like inference come into consideration. The schema and analysis is user-based and therefore inference oriented. Inference stands for inferring new information from already known data. Hence, a solution for an access control must include inference aspects as well. An integrated approach for access and inference control is required on a conceptual level. However, the main problem is not knowing which query or answer is problematic with the available data. This question primarily depends on the data provider. We summarize our work in three steps: Search the data security holes on a conceptual level, illustrated with the scenario of the GfK, give general indicators to address these problems and fix it with restriction or perturbation. The indicators can be precalculated and be integrated into an aggregation mechanism.

The remainder of this paper is organized as follows. In the next section, the structure of the multidimensional data model in general as well as specialities of the retail research at GfK is explained. In section 3, some aspects of inference and access problems are presented. Our indicator solution for these problems is proposed in section 4. The paper concludes with a summary and an outlook on future work.

1. MULTIDIMENSIONAL DATA MODEL

A data warehouse is a database integrating several data sources for analyses aspects. Data is often modelled on a conceptual level with a multidimensional model. We distinguish between a classification schema with the structure of a dimensions and the multidimensional schema which combines several dimensions and measures for a data cube.

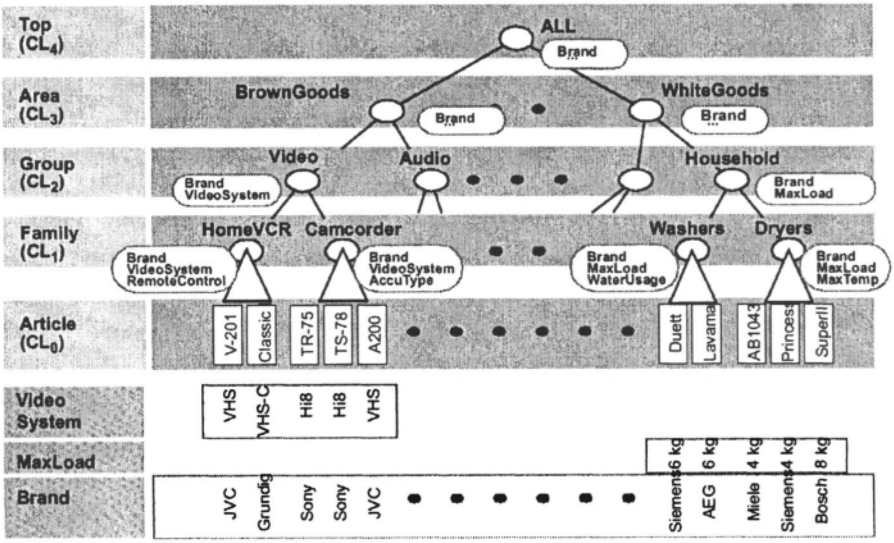

Figure 1: Classification schema and hierarchy of a product dimension

1.1 Classification Schemes and Classification Hierarchies

A classification schema defines the structure of the classification hierarchy. A classification schema *CS* is a partially ordered set of classification levels $(CL \cup \{Top\}; \rightarrow)$ where $CL = \{CL_0,...,CL_n\}$. Top is a generic element which is maximal with respect to "→", i.e. $CL_i \rightarrow Top$ for each $CL_i \in CL$. Figure 1 shows an example of a classification schema of the product dimension. The partial order "→" allows to place the classification levels in a directed acyclic graph (DAG). Top is generic in the sense that it is not modelled explicitly. CL_0 is the basic level of the DAG and called dimensional element.

Furthermore, each classification level SL_i of a classification schema *CS* has a finite domain $dom(SL_i)$. The instances of a classification level are classification nodes *CN* like HomeVCR for the classification level Family.

The classification nodes establish a tree structure. We assume the domains to be mutually disjoint and dom(Top) is {"ALL"}.

Moreover, the dimensions contain features like video system or brand. This information primarily depends on the instances of the dimensional elements, but can build up an inheritance hierarchy like in figure 1 and also depends on the classification nodes. Commonly, the features build another alternative classification hierarchy.

1.2 Multidimensional Data Structures

After having defined the dimensional data structure, this section deals with the formal definition of the multidimensional schema. A logical data model should clearly separate intension and extension of a data cube ([3]). Otherwise, logical schema design is not independent from the actual instances.

OLAP data cubes in general contain numerical measures, like sales or price. There are different types of measures with regard to their summarizability.

In fact, a multidimensional schema is a collection of possibly aggregated cube cells with a homogeneous structure. Homogeneous means, that all cube cells have the same granularity, i.e. aggregation level, and they contain the same measures. The multidimensional schema C is a structure C[G, M] where G = (G_1,...,G_n) is the granularity consisting of a set of classification nodes, i.e. $G \subseteq CN_{G1} \cup \ldots \cup CN_{Gd}$ such that for each $G_i, G_j \in G$: $G_i \neg\rightarrow G_j$ and M = (m_1,…,m_m) is a set of measures.

2. INFERENCE PROBLEMATIC

On the one hand, the multidimensional model supports the user in making decisions with its structure and operations based on the classification hierarchy. On the other hand, the user should not be allowed to ask all questions he wants to. In commercial systems, it is possible to deny the access for specific classification nodes or classification levels which can be done along with aggregation. But, a static access control on conceptual level is not enough, because the user is still able to bypass these reglementations through tricky queries. Therefore problems between the user's analysis focus and the data provider's limitations and dynamic aspects of querying in the multidimensional model have to be discussed.

2.1 One-Query-Inference

In this chapter the question is discussed which data can be inferred by one query, i.e. which single query leads to inference problems. The user himself has to search for the explicit data. The gist of a disclosure is to get sensitive data of an individual. In the multidimensional model, we call this a multidimensional related disclosure, because many dimensions are involved in one query. Besides, we distinguish between a refinement and a weighting disclosure.

Refinement

The refinement disclosure is an exact disclosure, i.e. a query specifies a query set. A query consists of measures and a combination of classification nodes from different dimensions. This is used to calculate a value with an aggregate function e.g. SUM over classification nodes. The result of a query is composed of the tuples of the query set.

Two examples of the refinement disclosure will be discussed. The first is called department store problem, the second query-set-smallness. The department store problem relates to a specific case of the market research company GfK. In this case a query set is specified by a query which results in a sum over the sales of a product in only two shops within a period. If this result is known by one of these two shops it can subtract its sales from the result of the query. The difference are the sales of the other shop. In this situation it is possible for one shop to determine the sales of its competitor. Let us explain it with the competitors A and B which are the only department stores in an specific area. The query 'Give me the sales of the department stores in Nuremberg of the TR-75 on the 15.04.1999' asked by A is a sensitive one, because the department store A can calculate the exact result with the subtraction of its own sales in this area and time.

The query-set-smallness can be seen as a generalization of the department store problem. It refers to each query that can be asked to the system and whose query set contains a low number of tuples. For example the query 'Give me the number of all sold TR-75 at the Media market in Nuremberg on the 15.04.1999' is a sensitive query. Its query set contains only one tuple and is therefore identifying.

Weighting

The weighting disclosure can be an exact or an approximative disclosure. Likewise a query set is used to calculate a value. But the feature of this approach is the domination of one or some tuples. They take a special position within the aggregation of the query set.

Related to our case study we found three types: the trade brand, the exclusive model and the weighting disclosure itself. Trade brands are brands whose producer and vendor is identical. Thus a trade brand is only distributed via channels of its own company. An example of a trade brand is Universum because the products of Universum are only sold by the vendor Quelle itself. A problem arises if e.g. the sales of a Universum VR in Nuremberg is known, then it is also known that the video recorder is sold by Quelle in Nuremberg. We call these problems trade brand disclosure.

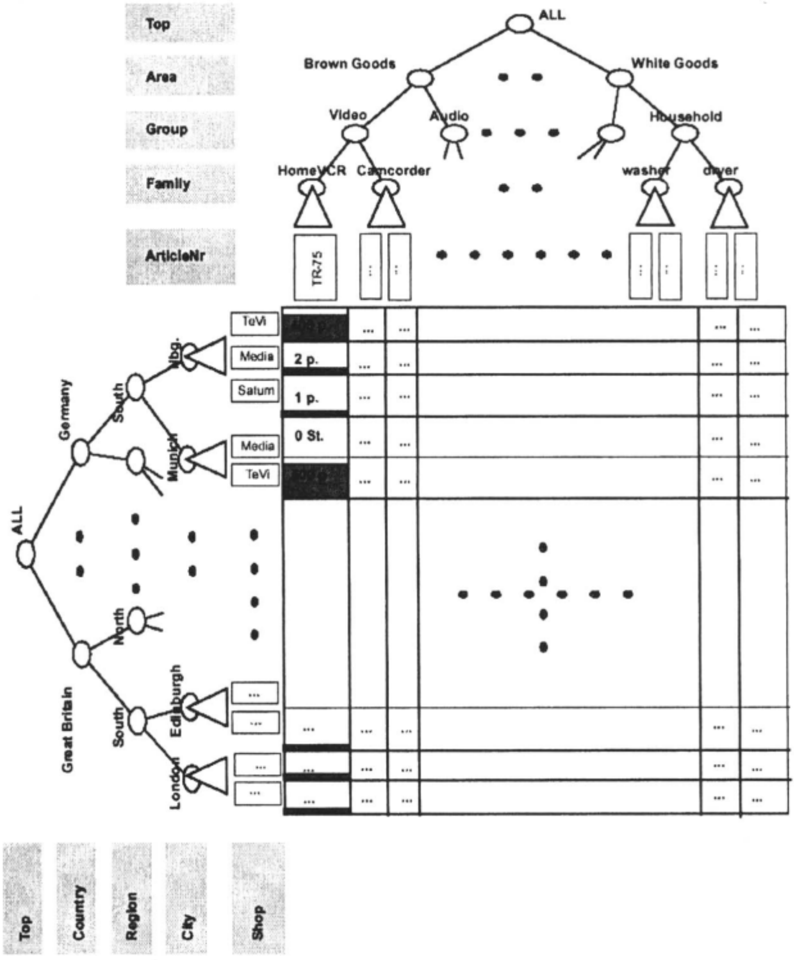

Figure 2: The weighting disclosure

Another issue is the exclusive model disclosure. Exclusive models are products a manufacturer makes only available for one vendor. In contrast to the trade brands the vendor can not be recognized by the vendor name but by the name of the product. An example of an exclusive model is a jubilee model, e.g. Sony delivers to Dixxon a special model with a designation only used for Dixxon. The inference of the exclusive models is similar to the trade brands. For both it is common that a query asking for sales of a trade brand or an exclusive model leads to a disclosure provided by a 100% domination of the tuples of the query set. A 100% domination is if all tuples of the query set contribute to the sensitive result that must not be known.

Sometimes, the domination is not a 100% domination but a k% domination. Logically a k% domination is the aggregation of some values of a combination of classification nodes divided by the aggregation over the same classification nodes of the whole query set. A more general type of disclosure is required as the trade brand or exclusive model disclosure. We call it weighting disclosure.

An example of the weighting disclosure is shown in figure 2. There is the product and shop dimension. Each dimension has its own classification hierarchies. The crosstable shows the number of sales. The query 'Give me the number of all sold TR-75 recorders in Nuremberg' or 'Give me the number of all sold TR-75 in South-Germany' results in a weighting disclosure. Because the first query consists of a three tuple query set (TeVi-, Media-, and Saturn), the TeVi result takes place a 400/403 = 99.3% domination. The second query leads to a 99.6% domination because of the domination of the two TeVi tuples within the query set. The weighting disclosure obviously does not lead to an exact disclosure, but it gives an impression of the sold products of a vendor if you know that domination.

Reducing the Result Set

Another trick to improve the result of a forbidden query is the parallel classification disclosure. It appears if two or more classification nodes of different parallel classifications are used to identify one dimensional element of one dimension without using the name of the dimensional element itself. In general the dimension element of a dimension is an unimportant information. Everybody knows that an electronic product like TR-75 exists. In all multidimensional data models the dimensional elements of all dimensions are well known. But, if there is a parallel classification of a dimension you can combine the classification nodes of different classification hierarchies. With this element it is not apparent for the system that you are on the forbidden level of the classification hierarchy.

In the example in figure 3, the product dimension has two parallel classifications: the hierarchy of the classification nodes and the feature classification. The product classification has a dimensional element which identifies exactly one result in the product dimension (extensional description [4]). The feature classification is not a dimensional classification, but an intensional description. Normally, it classifies not exactly one result, but you reduce the quantity of results. Both together are useful to determine one result in the product dimension, without using the name of the dimensional element. Universum color TV can be determined by an article number or as shown in figure 3 by using two intensional descriptions. The classification node CTV of the product classification and a special feature combination (68cm, stereo, 16:9, 100 Hertz, child prooflock) exists only for a Universum color TV which is equal to a specific product.

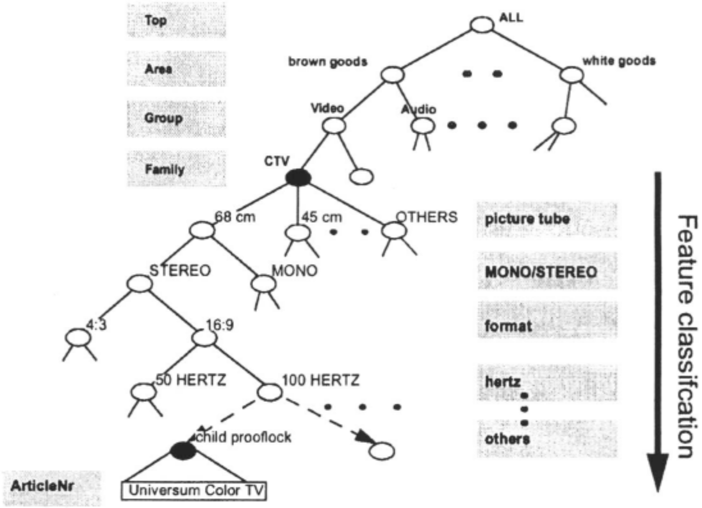

Figure 3: The parallel hierarchies ofthe product dimension

2.2 Multiple-Query-Inference

The one-query-inference isn't always successful because some static restrictions are able to suppress them. A more efficient and powerful way of getting sensitive data from a data warehouse is to go through a series of queries whose results are combined in a knowledge base (audit based approach). This memory is filled up until the search for sensitive data can be fulfilled. We distinguish two approaches: the intuitive and the theoretical approach.

Intuitive Approach

The intuitive approach is a combination of the parallel-classification- and the multidimensional-related disclosure. Intuitive means that the user is in the position to infer sensitive data without knowledge of mathematical retrieval methods. The user only relies on his knowledge of the market and combines it with the information of the reports to receive sensitive data. If the result isn't detailed enough, he uses the OLAP operations of the data warehouse to refine it stepwise. For example it is possible to get the sales of the trade brand Universum TV in Nuremberg by doing a drill down on the shop dimension to the classification node Nuremberg and on the product dimension to the classification node Color TV and then drill down the feature hierarchy. As the user is aware of the trade brand Universum TV and its characterizing features, he knows that he inferred the sales of the Universum TV of the Quelle shop in Nuremberg. The usage of OLAP operations in this approach bases on the knowledge of the user. They are only tools to infer sensitive data. This kind is used very often in the commercial area, because every manager has background knowledge about the competitors.

Theoretical Approach

A more complex, but efficient way is the theoretical approach. Some years ago, the area of scientific and statistical databases collected various methods of inference techniques and restriction of access. To give a short introduction, some ideas of revealing data are presented in the following.

A very efficient way of revealing particulars of an individual is shown by Denning at al [5]. They use a special characteristic formula called 'tracker' that is composed of the splitting of an forbidden, sensitive statistical query into smaller parts. Once found, a tracker is able to deduce the result of the sensitive query by involving the tracker and the response of a few answerable queries.

Rowe [6] does not try to gain sensitive information of an individual but statistics on a database which may also be sensitive. His approach 'diophantine inference' uses a special kind of diophantine equations that has to be solved to track down sensitive statistics. Palley and Simonoff [7] suggest an algorithm to yield a statistical disclosure of a confidential attribute. A synthetic database is created from the original statistical database. Then regression analysis is applied to it to estimate the hidden attribute. Delugach and Hinke [8] introduced an approach to automatically retrieve sensitive knowledge from a database based on 'Conceptual Graphs'.

All approaches have in common that the original usage was with unstructured data and not focussed on access ideas. The data warehouse environment and the multidimensional model contain the initial point for a new and broad data specific search.

3. INDICATORS FOR VISUALIZATION OF THE SECURITY HOLES

An overview of possible restrictions in statistical and scientific database area is given in [5]. Their methods mainly base on restriction of data and perturbation. A general solution is hard to present, because a generic perturbation contradicts the OLAP ideas and conflicts with economical operations like the distribution analysis, which requires data on a specific detail level. Otherwise a restriction of specific classification nodes, schemes or parts of the cube could be bypassed through other queries.

The main problem is not the restriction or the perturbation but to find a solution showing sensitive areas on the fly. We propose an indicator-based recognition on the conceptual level which can be often precalculated and used for an access control at runtime. Of course, only a small part of data is critical, i.e. not every query contains risks. But, queries on a detailed classification schema level should be handled with care.

An indicator is an inequation finding out that a query is sensitive and can compromise the system. If the inequation is true, it means that a disclosure is obtained; if it is false then the disclosure is not reached. These indicators can be precalculated in conjunction with an aggregation mechanism.

3.1 Indicator for Parallel Classification and Smallness

First we devote on the parallel classification and the smallness disclosure. Both use the same indicator, the smallness indicator:

$$k \le |K_G| \le N - K$$

N is the cardinality of the database and k is a threshold value. k depends on the user conditions and the limitations of the data provider. In case of the parallel classification disclosure $|R_G|$ is the cardinality of the used tuples of the dimension G. It recognizes whether a query reaches the classification level of the dimensional element of dimension G. A query is suppressed, if in all dimensions the level of the dimensional elements is reached. Otherwise $|R_G|$ is the cardinality of the query set R_G of the query characterized through its characteristic formula G. But, just as dangerous to find only a some tupels, too much tupels should be suppressed as well, because of the chance of building the compliment.

3.2 Department Store Indicator

The department store disclosure can't be avoided by the smallness indicator. We need a new indicator not to be dodged by a query. The department store indicator is only useful if the competitor, itself a member of the department stores, asks a department store query. Otherwise it should be true if a none member asks it. Of course, problems are still remaining, if other people do the search for these data.

$$\frac{SUM((product,\{x\})\cap(shop,\{a\}\cap\{C\}),sales)}{SUM((product,\{x\})\cap(shop,\{C\}\cap\{z\}),sales)} \ge s$$

This notation denotes a query for the sum of sales for a specific product

$$with\ \ s = 1 - \frac{SUM(\langle product, \{x\}\rangle \cap \langle shop, \{b\} \cap \{C\}\rangle, sales)}{SUM(\langle product, \{x\}\rangle \cap \langle shop, \{C\} \cap \{z\}\rangle, sales)}$$

and classification node in the shop dimension. A is a shop, and b its competitor, x is a product e.g. TR-75, z is a shoptype e.g. department store and C is a node of the classification hierarchy at which classification level the query is dangerous or not, e.g. C = 'Nuremberg'. The query in the denominator is the users query. With the instantiation of the variables the users query leads to the indicator s. So the system has to compute the query above - if the quotient is larger or equal than s, determined through the formula s, the query has to be suppressed.

3.3 Trade Brand or Exclusive Model Indicator

Trade brands imply a disclosure through the connection between the producer's and the vendor's and exclusive models between the product's and the vendor's name. The indicator below reveals a trade brand or an exclusive model disclosure.

x is a trade brand or an exclusive model, for C see above and a is the vendor of the trade brand or exclusive model. The query in the denominator is the users query. The query in the numerator has to be computed by the system and divided by the user's query. If the result is 1 the indicator signals that the user's query is a sensitive one.

3.4 Weighting Indicator

As mentioned in chapter 3 there exists a more general disclosure including the trade brand and exclusive model disclosure. The indicator for the weighting problem is described as follows:

$$\frac{SUM(\langle product, \{x\}\rangle \cap \langle shop, \{a\} \cap \{C\}\rangle, sales)}{SUM(\langle product, \{x\}\rangle \cap \langle shop, \{C\}\rangle, sales)} \geq s, s \in [0,1]$$

The difference to the trade brand or exclusive model indicator is that the quotient need not be 1 but over a certain threshold s to indicate a disclosure. The shop variable a need not be a shop of a trade brand or exclusive model but a user determined shop. Again, the query in the denominator is the users query.

4. SUMMARY

To find security holes is a fairly complex project, because both the user and the structure of the data model offer possibilities to achieve sensible data. In this paper, we presented some inference aspects in a market research company, to protect their data in a qualified way. Our indicator solution offers both, to the user a non static limitation and to the data provider the security not to disclose a secret. The next steps in our research will be practical tests with our indicator solution and the examination of data mining methods in this scenario.

REFERENCES

[1] Inmon, W.H.: Building the Data Warehouse, 2. edition. New York, Chichester, Brisbane, Toronto, Singapur: John Wiley & Sons, Inc., 1996

[2] Codd, E.F.; Codd, S.B.; Salley, C.T.: Providing OLAP (On-Line Analytical Processing) to User Analysts: An IT Mandate, White Paper, Arbor Software Cooporation, 1993

[3] Sapia, C.; Blaschka, M.; Höfling, G.; Dinter, B.: Finding Your Way through Multidimensional Data Models, in: 9th International Workshop on Database and Expert Systems Applications (DEXA'98 Workshop, Vienna, Austria, Aug. 24-28), 1998

[4] Lehner, W.; Albrecht, J.; Wedekind, H.: Multidimensional Normal Forms, in: 10th International Conference on Scientific and Statistical Data Management (SSDBM'98, Capri, Italy, July 1-3), 1998

[5] Denning, D.E., Denning, P.J., Schwartz, M.D.: The Tracker: A Threat to Statisical Database Security, ACM Transactions on Database Systems, 4(1), March 1979, p. 76-96

[6] Rowe, N.C.: Diophantine Inference on a Statitical Database, Infromation Processing Letters, 18, 1984, p. 25-31

[7] Palley, M.A., Simonoff, J.S.: The Use of Regression Methodology for the Compromise of Confidential Infromation in Statistical Databases, ACM Transactions on Database Systems, 12(4), December 1987, p. 593-608

[8] Delugach, H.S., Hinke, T.H.: Using Conceptual Graphs To Represent Database Inference Secruity Analysis, Journal Computing und Information Technology, 2(4), 1994, p. 291-307