Available online at www.sciencedirect.com
SCIENCE DIRECT®
Omega 32 (2004) 87–95
omega
The International Journal of Management Science
www.elsevier.com/locate/dsw

A selective data retention approach in massive databases

Orly Kalfus, Boaz Ronen, Israel Spiegler*

Technology and Information Systems Department, Faculty of Management, The Leon Recanati Graduate School of Business Administration, Tel Aviv University, Tel Aviv 69978, Israel

Received 20 February 2002; accepted 26 September 2003

Abstract

Exponentially growing databases have been tackled on two basic fronts: technological and methodological. Technology offered solution in storage capacity, processing power, and access speed. Among the methodologies are indexing, views, data mining, and temporal databases, and combinations of technology and methodology come in the form of data warehousing, all designed to get the most out of and best handle mounting and complex databases. The basic premise that underlines those approaches is to *store everything*. We challenge that premise suggesting a selective retention approach for operational data thus curtailing the size of databases and warehouses without losing content and information value. A model and methodology for selective data retention are introduced. The model, using cost/benefit analysis, allows assessing data elements currently stored in the database as well as providing a retention policy regarding current and prospective data. An example case study on commercial data illustrates the model and concepts of such method.

Keywords: Databases; Data warehousing; Data mining; Information as inventory; Cost/benefit analysis

1. Introduction

The growing volume of data in very large databases creates both a problem and opportunity that results in new types of systems and analytic techniques. Those large databases have been embraced on two basic fronts: technological and methodological. Technology offered solutions in storage capacity, processing power, and access speed. Methodologies such as indexing [1], views [2–4] temporal databases [5,6], and data warehousing [7]—a combination of technology and methodology, all have been designed to get the most of and handle mounting and complex databases.

The problem can be traced back to the early, heady days of the information revolution, characterized by rapid development and mushrooming of systems demanding more and more hardware and storage facilities. This information explosion, as it came to be known, had all the symptoms of untrammeled growth: systems were out of control, databases "polluted", and right information scarcely available for decision making.

Data is usually referred to as the raw materials from which meaningful information is produced. Both are then used to gain insight and generate knowledge [8]. As such, data must be regularly examined for their contribution and value to decision makers, discarding aged and valueless facts. New data, either about to enter a firm's database or data warehouse must be regularly assessed before use. Continuing with the raw material analogy, we advocate treating data as inventory in a production system, in line with concepts like *just in time* employed in operations management [9,10].

Data and storage evaluation techniques and methods for information systems are scarce. Two approaches have been taken to tackle the problem of over-abundance of data:

- *Technologies*: to improve size, speed, and processing power, and increase and extend the capacity of storage media.
- *Methodologies*: to enhance query efficiency and performance, usually at additional storage cost and processing power.

* Corresponding author. Tel.: +972-3-640-7488.
E-mail address: spiegler@post.tau.ac.il (I. Spiegler).

0305-0483/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.omega.2003.09.015

Technologies and methodologies developed have improved processing power, condensing and compacting data storage, enhanced representation of data, introduced indexing, view materializing, temporal techniques, and others [11,12]. Data warehousing, a combination of technology and methodology, is a good manifestation of such solutions to large databases. Concurrently, data mining (DM) methods are developed to extract knowledge from volumes of data. Those methods generally take the database as given without tempering with source data or structure but rather applying models to extract new information and knowledge.

But, storage technology can hardly provide a long-term solution to persistent accumulation of data by individuals and organizations. Whatever its size or capacity, the "container" fills up at a rate that seems to match and surpass any technological developments. The history of data management from file organization to hierarchical and relational structures, data warehousing, and data mining, is a point to that effect.

The mountains of data stockpiled in databases are usually operational data. Some or all data from operational databases is transferred into a data warehouse to which DM and On Line Analytical Processing (OLAP) [13] can be applied. This is a present phase in data management the objective of which is to identify valid, novel, useful patterns and associations in existing data in order to gain insights that add to an organizations knowledge [14,15], or "to extract higher level information from an abundance of raw data" [16]. This process is also called knowledge discovery in databases (KDD) [17,18].

While not attempting to reduce the *size* of the database, those techniques and methodologies aim at extracting and distilling "new" information from the stored operational data. Many techniques and models have been developed in DM, among them: classification and clustering [19], regression [18], association [16], statistical methods and neural networks [20,21]. For an extended survey of DM models see [18]. View materialization [3,4] and indexing techniques [1] are other techniques used to manage large databases without interfering or changing the data content or structure.

Still, the basic premise that underlines all those approaches is to *store everything*. In a sense such approach is indeed lack of data policy avoiding or deferring decisions on storage. We argue that making a "to store or not to store" decision on data components, either upon entry or while residing in the organizations database, may reduce the dimensions of the problem and yield a more effective data management in a firm.

We thus challenge the premise that all data collected by an information system are essential and hence must be stored in its database. Employing a "only the data that is fit to store" approach where organizations *remember selectively* rather than *remember everything* is the argument to produce direct cost savings and improve efficient data use. Nevertheless, selective data retention does not rule out or replace data mining but rather complements it.

Few studies question the basic premise that all data are necessary and hence must be retained. We do not propose that collected data be physically expunged, rather be removed from the database to lower level storage archives, if necessary. The idea is to treat the data as inventory in a production system [22] where data are constantly assessed and that new data about to enter the system carefully evaluated as part of a quantitative and qualitative appraisal of information systems.

Our approach is founded on the idea of information as inventory introduced by Ronen and Spiegler [22]. The idea draws from the work in process (WIP) concept used in operations and inventory management [23,24]. Following are some of the ills of information in process (IIP) pointed out by the authors:

1. Information in process lengthens the *cycle time* needed for decision making.
2. The quality of information produced from a voluminous database is likely to be lower than one produced from a more compact database since the system must process unneeded or irrelevant facts, rather than focusing on relevant information.
3. Age of data is often ignored in making storage decisions. Older facts take up storage space, impede access to needed information, and may mislead in understanding reality.
4. Management, control, and maintenance of a system are harder when much data and complex storage media need to be handled.
5. Automatic and indiscriminate data entry is costly and time consuming.

Accumulating data indiscriminately stems from the notion that storage cost is low. It is based on the fallacious idea of a decreasing unit cost per gigabyte of data when the real key issues are total organizational cost and the fact that IIP increases at a faster rate than unit cost [25]. Goldratt [26] and Kaplan [27] promote the same idea in the operations management.

Among the elements contributing to the evils of IIP are: the use of a local unit cost for data storage, the tendency to introduce new technology, and focusing on *efficiency* rather than effectiveness of an information system [12]. All these ills are amplified by the increase in the actual cost of storage and maintenance as well as software and human costs needed to manage the database and data warehouse.

Hence, it becomes necessary to evaluate the retention in storage of specific data fields. The objective of this paper is to show how such evaluation may be done, and to test and validate the concepts of selective retention of data.

We outline a model and a methodology for assessing the data stored in a firm's database. Our basic unit of analysis is a *field* which represents a single attribute of data. The term *data element*, often used as synonymous, really applies to an information system as a whole. The model establishes

criteria for deciding whether or not a field is to be retained in the database. The same analysis can be applied to data elements in an entire system. The model uses cost/benefit techniques on the fields of a database. An example case study of a commercial file is used to illustrate the model and concepts described. Results of the preliminary study indicate that elimination of some data hardly impaired the value, quality, and contribution of information to decision making.

The paper has the following parts. Following the Introduction that reviews the basic concepts related to this work, Section 2 describes the model, data element profile, criteria, and range of data policies for database management. A selection methodology for volume handling in databases and warehousing is also outlined. Section 3 presents a case study to illustrate cost/benefit analysis of selective data retention on a commercial file. The final section summarizes the study and suggests directions for further research.

2. The model

The proposed solution of our study consists of two parts: a model to define the variables, costs and benefits relating the data resource, and a methodology for drafting policy and making decisions about selective retention of data.

One way to embark upon large volume of data is to classify and categorize them. A simplified way is the Pareto 80:20 principle, extended into a three-categories model, which claims that inventory or data elements vary in relative importance [24]. Thus about 80% of the system's functions are controlled by 20% of the data population. This principle of categorization is applicable also to handling data, the implication being that management should focus on the more important fields, which constitute a small portion of the system's data content.

As mentioned, our basic unit is the data *field*—a unit that can be measured and tested either in an objective or subjective manner. The model is generic in the sense that the basic unit of analysis may also be a record, file, or even a data set in a large system. A *data profile* is a set of independent features or characteristics of any such data element. The profile will generally contain parameters of two types: (1) objective parameters, and (2) subjective parameters such as relative importance that are obtained in an interview or evaluation.

Using relative importance of a field can be prioritized according to parameters such as cost of the element, how critical it is, the cost of lacking it, the effort to collect it, and the like. The formal definition of variables and parameters used in the model follow.

2.1. Data profile

Let $\{x_1,\ldots,x_5;\ q_j\}$ be field variables as below, and $j = 1,\ldots,J$ where J is the number of processes. Table 1 gives a set of independent variables that are estimated for a field to determine its importance. The parameters that make up the field profile are shown in Table 1.

2.2. Importance of a field

The relative importance of a field in a work process is given by

$$h = \sum_{i=1}^{4} x_i w_i + \sum_{j=1}^{J} q_j w_j,$$

where h is the overall importance index of a field, x_i the characteristics of a field, independent and between 0 and 1, w_i the relative importance (subjective) of each of the characteristics, q the how critical the field is to a work process, w_j the perceived importance (subjective) of a work process j, and J the processes relevant to a system (or subsystem) under study.

The first term defines the importance as it indicates factual measures such as field usage (x_1), relevance (x_2), currency (x_3), and availability (x_4) multiplied by a subjective weight w_i. The second term describes the perceived importance of the data element; it derives from the perceived importance of work processes and how critical the field is to a process. Since both w_i and w_j are subjectively estimated the overall importance h is subjective too and on the same scale. Parameters estimation is done by session of the analyst with decision-makers and database administrator (DBA). Mandatory considerations such as legal requirements to store a field

Table 1
Data field's profile

Parameter	Meaning and use	Need and values
Field used (x_1)	Is the field used as input to processing programs	0 = No, 1 = Yes % of use in system
Relevance to population (x_2)	Is the field relevant to other elements or occurrences	0 = No, 1 = Yes % of use in system
Updating relevant population (x_3)	Are there data in the defined range updated by the field	0 = No, 1 = Yes
Field raw or result of operation (x_4)	Is the field a source data or produced by some routine	0 = Yes, 1 = No
Field supplied from other sources (x_5)	Is the field coming from other (outside) sources	0 = Yes, 1 = No
Critical for work process $j(q)$	Is the field critical to running of a process in system	0 = No, 1 = Yes

(e.g., bank transactions), medical factors, or archival needs certainly overrule such considerations.

2.3. Data policies

There are five alternative selective data retention policies with regard to a field:

(1) Retain the field in the system.
(2) Retain the field but change its structure. This alternative deals with fields that are relevant only to part of the overall record population. Isolating the relevant part may result in substantial storage savings. For example, consider a file of 100,000 records, where each record has 100 bytes. The file may have a primary key field of 15 bytes and another field of 10 bytes relevant to only 5% of the population. Pulling out this field from the file means the system will require

$$(15 + 10) \times (0.05 \times 100,000) = 125,000 \text{ bytes},$$

instead of

$$100,000 \times 10 = 1,000,000 \text{ bytes},$$

a storage saving of 87.5%. Thus, a change of structure should be considered. Normally, including a field in a file relevant to only 5% of the population is a design error probably introduced at file design time. Still, such field may be of critical importance to the database, e.g., a medical symptom of fatal potential, or an attribute of a highly valuable customer, at which point no removal or restructure will be done.
(3) Remove the field and produce it upon demand. The benefits here are storage saving and reduction in maintenance. The cost (a negative benefit) is to produce the field rather then retrieving it, i.e., increasing response time. Any removal of data is not related to historical fields that are kept for legal or contractual purposes.
(4) Remove the field from the system and supply it from another source upon demand. The benefits here are storage savings and reduction in maintenance. Increasing response time is a negative benefit, i.e., a cost.
(5) Remove the field from the system and do not supply when demanded. The benefits and costs are similar to those of alternative (4).

Cost of each policy is defined formally as follows:

C_1, the cost of retaining a field in the system is

$$C_1 = C_s + C_m + C_u,$$

where C_s is the cost of field storage, C_m the cost of regular updates including software maintenance, and C_u the cost of user's time.

C_2, the cost of retaining the field after a change in data structure. The cost of this policy is cost of conversion + cost of supplying field from new structure −cost of continuing to supply the field from the old structure.

$$C_2 = C_c + C_p + C_r - C_1,$$

where C_c is the cost of converting current structure to another structure, C_p the cost of converting a program to new data structure, and C_r the one time cost to hold the field in new data structure.

C_3, the cost of producing the field, defined as

$$C_3 = C_f + C_y,$$

where C_f is the cost of removing a data element from the system, $C_f = C_c + C_p$, and C_y the development cost (hardware, software, and people).

C_4, the cost of supplying the removed field from another source is

$$C_4 = C_f + C_b,$$

where C_b is the cost of collecting a data element from an external source.

C_5, the cost of not having a field when demanded, is defined as

$$C_5 = C_f + C_a,$$

where C_a is the cost of lacking a field (negative benefit).

Both C_4 and C_5 are manifestations of the x_5 profile parameter defined in Table 1.

2.4. Benefits

There are two aspects to estimating benefits (B): actual cost saving, and improving the information quality.

Cost saving: The reduction of cost as a result of taking an alternative is expressed as $P_s \times B_s$ where B_s is the saving resulting from any given alternative O and P_s is monetary savings per unit of storage space.

Savings come from hardware, software, and mainly from labor intensive cost of maintenance of the database. There are also savings in updating and in response time but these are negligible.

Information quality (V): Three factors contribute to improving the information quality in a system:

1. Improving response time (T). The information usage increases as response time decreases. In terms, if T is the time elapsing from the request for information (t_0) until its supply (t), then $T = t - t_0$.
2. Increasing the ratio of important information. That is, apply the relative weight of a category to its importance. We denote
   H—improvement in the average level of importance of a category, where k is the number of units in the category.

$$H = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{j=1}^{4} X_{ik} w_{ik} + \sum_{j=1}^{J} q_{ik} w_j \right).$$

3. Raising the quality of the stored data to produce information. Discussion on the quality of information is given below.

Table 2
Data categories and corresponding policy

Data category	A	B	C
Importance level	High	Medium	Low
Field criteria	If field is: 1. in use x_1 2. relevant x_2 3. updated x_3 4. raw data x_4 5. critical to process x_5	Field belongs neither to A nor to C	If field is not used ($x_1 = 0$) AND not updated ($x_3 = 0$) OR the field is result of operation ($x_4 = 0$)
Policy action	Retain field in system	Alternatives: 1. retain (A) 2. retain, and change data structure 3. remove, produce if and when needed 4. remove, supply from external source 5. remove, do not supply	Remove field from system if: cost to retain > cost of removal. (for criterion no. 2, produce if no effect on response time).

2.5. Categories and policy

We define three categories of data importance for selective criteria:

- A—highest degree of importance,
- B—intermediate, and
- C—lowest importance,

i.e., $q\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. The classification criteria and corresponding policy action are shown in Table 2. Decisions relevant to fields in the system are $D = \{$retain, remove, structure-change$\}$. These decisions are illustrated in the example case study (Section 3).

2.6. Selecting methodology

Following the description of the model and cost/benefit analysis, we outline a preprocessing data methodology that consists of steps to help structure policy decisions and possible action to manage a database or data warehouse. The steps are:

1. *Establish organizational and information system goals*. This is the basis for determining the importance of data, processes, and the type and quality of information the system renders to its user community.
2. *Identify boundaries and data population*. Boundaries determine the scope of a system under study in terms of sub-systems, files, records, and data elements relevant to the desired process or operation.
3. *Divide data into categories*. A category is one or more data elements having a common base in terms of physical or logical structure, or having the same *profile*. Determining categories can be based on age, volatility, level of processing, collected vs. generated data, and others.
4. *Define data profile*. We defined above a field as a distinct component that can be measured and tested as a unit. A *profile* is a set of independent features, objective or subjective of the field. The profile is general and equally applicable to different levels of data: field, record, file, database, or information system.
5. *List decision and work processes*. The list of processes derives from the organizational and IS goals. Relevant decisions and policy are particular for each case.
6. *Estimate parameters for data elements*. This step involves obtaining parameter values, as they appear in Table 1, for each relevant field. The idea is to translate the values into common monetary terms, a process that is done objectively or subjectively.
7. *Perform cost/benefit analysis*. The cost/benefit analysis is performed for each of the five alternatives for handling a data element (see Table 2).
8. *Operational decision*. This step generates a policy decision, based on the cost and benefit analysis, regarding action to be taken for data retention.
9. *Overall cost/benefit calculation*. After completing steps 1–8 for all fields in the study, a cost/benefit analysis is done for the entire data set retained in the system.
10. *Follow up*. The last step establishes a follow-up routine for data policy by regularly evaluating the decisions taken. This includes applying the methodology on a regular basis or at new data warehouse generation takes place.

2.7. Quality of information

Our study included also aspects of the quality and value of information. These are briefly summarized now. We distinguish between data quality and information quality. Data quality consists of aspects such as accuracy, currency,

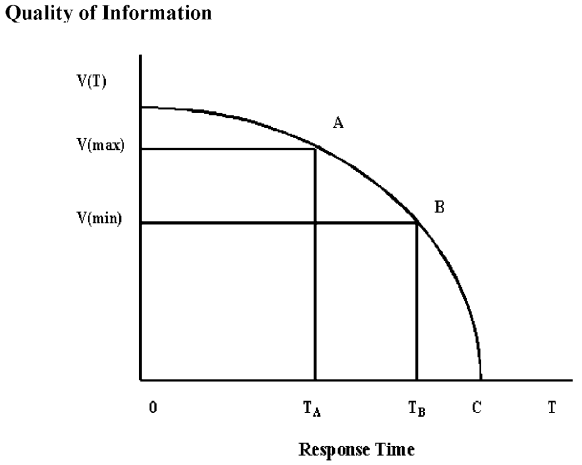

Fig. 1. Information quality vs. response time.

frequency, scope, and mode of presentation. While in many cases data quality and information quality are positively related, this may not be the case when data is over detailed or too frequent. A subjective assessment by users determines how much a given level of quality is desired or worth.

Information quality has to do with the product generated from the database. Several aspects influence the quality of information, one of which is response time. We assume that the quality of information decreases as response time rises. It views information as becoming less relevant over time, as shown in Fig. 1. Maximum information quality ($V_{\max}$) is independent of the change in response time ($\mathrm{OT_A}$). From point A, the quality of information decreases exponentially to point B which is the minimum information quality, $V_{\min}$, the system should supply. Beyond that point and to C information quality is negative.

The information quality can be measured in one of two ways: first, by using concepts such as entropy and normative value of information [28]; and second, by applying a relative measure whereby the information supplied on time gets the value 1. Thus, the information quality at $V_{\max} = 1$, and the numbers on the $\mathrm{OV_{max}}$ axis are between 0–1. The rate of decline can be estimated by asking the user, the supplier of the information, or a decision maker to attach a monetary value to information loss.

On the basis of these assumptions and the graphical representation in Fig. 1, the relation between information quality and response time can be expressed as

$$V(T) = 1 - \left[\frac{1}{L_1} e^{(T-T_A)/L_2}\right] \Big/ 100,$$

where L_1 and L_2 are constants affecting the curve, and dividing by 100 sets the values of V in the range of 0–1.

The importance of a field and the information quality are positively related. And surely, the higher the ratio of type A component, the higher the quality of the information.

Improvement of information quality are directly linked to update level, accuracy, reliability, level of detail, and for of data storage and representation. It is a subjective measure that combines the elements mentioned above that should be express in monetary terms. This is beyond the scope of our present study.

3. Case study

To illustrate and test the model described above we applied our approach to a data set taken from a large insurance agency which has been in business for over 25 years, selling most types of life insurance policies to firms and individuals. The agency provides a wide range of services to its customers from policy sales through termination. The agency's information system, which was developed in-house, has been in operation for over 15 years [29].

We selected one file from the agency's client database, and took it through all the steps of the methodology, including organizational goals, processes, boundaries, data profile, categories and cost/benefit analysis. The file comes from a relatively stable database with low volatility or structural changes. It contains typical facts such as customer name, address, telephone number, and other insurance data. For our study, we note its following characteristics:

Data	Detail
Number of fields	59
Length of record	246 bytes
Age of record	10 years
Type of data	Mainly raw data
Number of occurrences (randomly selected from file)	1000

The field profile includes the variables detailed in Table 1. Both the analysts and users were asked to determine the variables by indicating a yes/no answer (either 0 or 1). For x_3—does the field update other data—we used an update threshold of 70% of the time as being relevant.

Six work processes were identified, and were given equal importance. Thus, the range of values for perceived importance of the field is between 0 and 6, which is the number of different process possibilities. The processes are

Process	Description
w_1	Insurance design for client
w_2	Client consulting
w_3	Insurance sale
w_4	Accounts receivable
w_5	Information services
w_6	software maintenance

Table 3
Classification of fields in case study

	Category A	Category B	Category C	Total
Number of fields	31	17	11	59
% of total	52	29	19	100
Length in bytes	125	84	37	246
% of total	51	34	15	100

The full range of measurements on data elements is not shown.

The categories are defined according to the criteria set in Table 2. The classification of data elements is shown in Table 3.

Table 3 depicts the number of fields in each category and their corresponding length in bytes as a measure of space. This initial analysis shows that category C has 19% of the fields and 15% of the record space. Since members of category C are subject to removal from the file, this space may be saved. On the other hand, Category A takes 52% of the fields and 51% of the space, which indicates high importance that should be retained.

3.1. Cost/benefit analysis

The cost/benefit of three possible data decisions: (1) retain in the system, (2) remove from the system, and (3) remove with structural change, produced the following results:

	Retain	Remove	Structure change
Fields	41	14	4
Bytes	175	64	7

1. *Retain in the system*: Data to remain in the system are category A, those parts of category B which have no other source and are needed for automatic processing by an external user, and data impossible or impractical to make a structural change. The following results show that 71% of the record (175/246 bytes) will remain in the system.

2. *Remove from the system*: Data to leave the system are category C and part of category B, which amounts to 14 fields or 64 bytes (26% of the file space). Removal costs include physical conversion cost, cost of program updates, and cost of user program compilation.

3. *Structural change*: Structural changes were needed in four fields of the file. An example of a structure chance is a "flag" field of 1 byte (having the value 0 or 1) that appears in only 2.8% of the data set. The field was replaced by a small file, which contained only the flagged records. Using the sample of 1000 records, we get

$$1 \times 1000 \times 3\% = 30 \text{ bytes flagged, where}$$

$$30 \times 13(\text{key length}) = 390 \text{ total bytes needed.}$$

Compared to the 1000 used in the original file, saving of 61% in storage with no information loss. For the other three fields, structural changes led to a saving of 87.5% of space. In a large data warehouse such saving becomes significant both in storage and processing time.

Overall retention of sample data set, after removal and structural changes, was 71% of the file space (174,660 bytes instead of original 246,000 bytes). Fig. 2 shows the results before and after the applying our model. Note, that in the new structure, category A amounts to 71% of the data 51% before the change. Category B will have 50 bytes (29%) and no category C remained after applying a selective retention methodology.

4. Summary and conclusions

A model for selective retention of data in massive databases was presented. It is suggested as an alternative to the *store everything* practice found in many databases and data warehouses resulting exponential growth in storage space. The model defines a data element (e.g. field) profile and employs cost/benefit analysis to assess data currently stored in a database. Motivation for the selective retention approach come from the idea of Information as Inventory which advocates applying operational management concepts to databases and information systems.

Applying the model and cost/benefit analysis to a commercial data file suggests two notable results. First, selective retention methodology can bring about 30% of storage saving in databases or data warehouses. Second, the analysis results in a major shift in importance as reflected by data categories. Following application of our model, category A (most important) increased from 51% to 71%, eliminating category C (least important) altogether.

The model may be useful in applications other than storage savings and data management. It may contribute in data cleansing and feature selection activities of data mining projects. Those are routine processes done on databases and data warehouses to get better information and knowledge. And the model and methodology may offer a framework by which managers can begin to draft selective data policies and alternatives to store everything practice in organizations.

Further studies are certainly needed, with larger and different databases employing the selective retention approach. Another study is a follow up on removed fields from a database in terms of types I and II errors taking a different angle to the same subject.

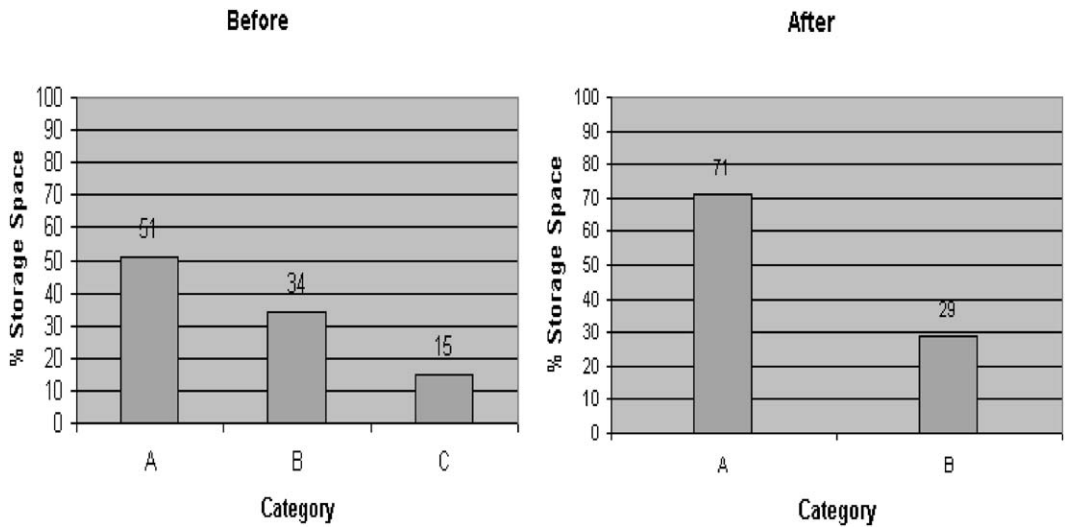

Fig. 2. Data categories—before and after model application.

References

[1] Cox K. A unified approach to indexing and retrieval of information. Proceedings of Conference on Technical Communications at the Great Divide, Alberta, Canada, 1994. p. 176–81.

[2] Agrawal D, El Abbadi A, Singh A, Yurek T. Efficient view maintenance at data warehouses. Proceedings of the ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, 1997. p. 417–27.

[3] Jones MC, Rundensteiner EA. View materialization techniques for complex hierarchical objects. Proceedings of the Sixth International Conference on Information and Knowledge Management, Las Vegas, Nevada, 1997. p. 222–9.

[4] Quass D, Widom J. On-line warehouse view maintenance. Proceedings of the ACM SIGMOD, International Conference on Management of Data, Tucson, Arizona, 1997. p. 393–404.

[5] Elmasri R, Kouramajian V, Fernando S. Temporal database modeling: an object-oriented approach. Proceedings of the Second International Conference on Information and Knowledge Management, Washington, DC, 1993. p. 574–85.

[6] Gadia SK, Yeung CS. A generalized model for a relational temporal database. Proceedings of the Conference on Management of Data, Chicago, IL, 1988. p. 251–9.

[7] Inmon WH. Building the data warehouse. 2nd ed. New York: Wiley; 1996.

[8] Spiegler I. Knowledge management: a new idea or a recycled concept. Communications of the AIS 2000;(14):1–24.

[9] Schonberger RJ. Japanese manufacturing techniques. New York: Free Press; 1982.

[10] Suzaki K. The new manufacturing challenge. New York: The Free Press; 1987.

[11] Bell T, Witten LH, Cleary JG. Modeling for text compression. ACM Computing Surveys 1989;21(4):557–91.

[12] Ein-Dor P, Carl RJ. Information system management: analytic tools and techniques. New York: Elsevier Science Inc.; 1985.

[13] Dinter B, Sapia C, Hyfling G, Blaschka M. The OLAP market: state of the art and research issues. Proceedings of ACM Workshop on Data Warehousing and OLAP, Washington, DC, 1998. p. 22–7.

[14] Chung HM, Gray P. Data mining. Journal of Management Information Systems 1999;16(1):11–6.

[15] Ramakrishnan N, Graama AY. Data mining: from serendipity to science. IEEE Computer, 1999;32(8):34–7.

[16] Aumann Y, Lindell Y. A statistical theory of quantitative association rules. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, August 15–18, 1999. p. 261–70.

[17] Communications of the ACM. Knowledge Discovery 1999;42(11) (Special Issue).

[18] Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining. Cambridge, MA: AAAI, MIT Press; 1996.

[19] Ganti V, Gehrke J, Ramakrishnan R. Mining very large databases. IEEE Computer, 1999;38–45.

[20] Adriaans P, Zantige D. Data mining. Reading, MA: Addison-Wesley; 1996.

[21] Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. Discovering data mining: from concept to implementation. Englewood Cliffs, NJ: Prentice-Hall (IBM); 1998.

[22] Ronen B, Spiegler I. Information as inventory: a new conceptual view. Information & Management 1991;21: 239–47.

[23] Ronen B, Palley MA. A topology of financial versus manufacturing management information systems. Human Systems Management 1987;7(4):291–8.

[24] Ronen B, Pass S. Manufacturing management information systems requires simplification. Industrial Engineering 1992;24(2):50–3.

[25] Eden Y, Ronen B. Service organization costing: a synchronized manufacturing approach. Industrial Management 1990; September/October:24–6.
[26] Goldratt EM. The haystack syndrome. Groton-on-Hudson. New York: North River Press; 1991.
[27] Kaplan RS. Relevance lost—the rise and fall of cost accounting. Boston, MA: Harvard Business Press; 1987.
[28] Ahituv N. A systematic approach toward assessing the value of an information system. MIS Quarterly 1980;4(4): 61–75.
[29] Eden Y, Ronen B, Spiegler I. Storing too much information in the insurance industry. International Journal of Computer Applications in Technology 1996;9(2–3): 144–50.