Overview of some Contribution for Business Intelligence research

Veit Köppen

Freie Universität Berlin
Institut für Produktion, Wirtschaftsinformatik und OR
Garystr. 21, 14195 Berlin
koeppen@wiwiss.fu-berlin.de

Summary. In the last years Prof. Lenz has contributed a lot of work on Business Intelligence. This articles tries to give a short overview of selected papers. Since 1975 Hans-J. Lenz published more than 350 scientific books, papers and reports. A brief overview on the work on Business Intelligence is given in this article. Corresponding to the structure of this book the following articles are outlined.

1 Introduction

In the last coupling years Prof. Lenz dedicated much of his research to Business Intelligence. Some of his work is briefly summarized in this article. A definition and explanation of Business Intelligence is given in (**?**). In the following the main objects of research are presented. The architecture of Databases including statistical databases is presented in the next section. Work corresponding to data warehouse architecture and the fusion of data to obtain a "good" data warehouse is presented in the following section. As a new trend mobile OLAP is also included in this research area.

Another section deals with Uncertainty within business's. This may arose due to predicting the future or work on data with measurement errors. Another possibility is the inclusion of incomplete or inaccuracy data. That the controller has to make his decisions upon such error-based data is normal life. Decision assistance and helping decision makers is another section in the research area of Prof. Lenz.

Applications where business Intelligence is used are also proposed in several papers and work. Here the work on knowledge management and planning travels is given.

Statistics play a very important role in Business Intelligence and the work of Prof. Lenz. But this is only one reason why some of his work regarding statistical research is presented in a last but not least section.

2 Architectures for Statistical and Scientific Databases

A lot of publications exist on architecture of databases. But using aggregate functions like SUM, AVG and $MEAN$ demand also data of the data. The M^3-architecture utilizes micro-, macro- and meta data. Meta data is viewed as data upon micro- and macro-data. Whereas micro- and macro-data are fixed formatted meta data are highly cross-referenced and are extremely heterogenous. In meta data special attentions has to be brought to data integrity and confidentiality. The focus on the conceptual design is given in (1), (2), (3), (4), (5), (6), (7).

3 Working in a Data Warehouse

In the following section the work especially dedicated to data warehousing is presented. Restrictions that are due to the special characteristics of data warehouses are investigated and contributions to these characteristics is given.

3.1 OLAP

On-Line Analytical Processing distinguishes from On-line Transaction Processing mainly due queries that are build upon $SELECT$ clauses. But also the intensive use of aggregate functions distinguishes both methods.

In (8) the summarizability of OLAP and Statistical Databases is investigated. To avoid erroneous conclusions and decisions it is important follow the conditions for summarizability. A framework is introduced, where the context in which statistical objects are defined is precisely specified.

Aggregation functions are a class of generic functions which must be usable in any database application. In which case the aggregation functions can be correctly applied on macrodata is characterized (9). The macrodata are build up the data cube and are computed on the microdata.

A theory of aggregation functions, OLTP-OLAP transformations, and of the data cube is developed in (10).The OLTP-OLAP specification frames is an architecture for OLTP-OLAP applications that supports sound and correct querying.

3.2 Data Fusion

Data in Data Warehouses are merged by a huge amount of data resources. But the fusion of heterogenous sources is not a trivial task if no identifying item within all data sources exists.

The semantic discrepancies between data sources is researched in (11). Object identification is an essential method for the task of data integration. The use of similarities, rules or classification criteria with different statistical

or data mining techniques can be applied to classify pairs of records from different sources in order to link them or not.

The problem of object identification also occurs in census data, where several administrative registers are merged. For instance this is true for German census because no unique identification number exists in the registers. A twostep procedure consisting of a preselection technique of pairs and the classification of "matched" or "not matched" is presented in (12).

Another approach deals with numeric data. To avoid semantic incoherency with respect to the knowledge at hand the input of numeric data into databases needs a careful screening. In reality this is done by knowledge like balance or behavioral equations or definitions. (**?**) present validation rules, also known as edits that can represent the knowledge and improve semantic consistency.

3.3 mobile OLAP

Ubiquitous access has become a critical success factor in most enterprizes. Wireless networks are expected to host data warehousing applications. But due to network transfer and memory limitations

Time critical decision-making is nowadays mostly done by multidimensional data usage in an wireless network environment. But additional complexity is introduced by the shortcomings of both systems and due to mobile devices operating within their proximity. (13) deals with the efficient dissemination of multidimensional data into wireless networks. A new family of scheduling algorithms, which simultaneously exploits various characteristics both of OLAP data and wireless networks, is introduced.

A hybrid scheduling algorithm, explicitly suited for on demand delivery of aggregated data in wireless networks is proposed in (14). By selecting between highly compressed data structures, the size of data transmitted into the network is minimized, however without any loss of semantic information. A reduction of access time and device energy consumption, but also a minimization of generated traffic is achieved with this algorithm.

3.4 Index Structures for Data Warehouses

Short response times are essential for on-line decision support. Common approaches to reach this goal in read-mostly environments are the precomputation of materialized views and the use of index structures.

In most research only one topic is treated. (15) examines a possible combination of both techniques. The R^*-tree is used as an example of a multidimensional index structure. Aggregated data is stored in the inner nodes of the index structure in addition to the references to the successor-nodes.

Materialized aggregates in the inner nodes of such index structures are used to speed up range queries on aggregates. Existing models are extended to take account of aggregated data in (16). A new generic performance model to estimate the Performance of Index Structures with and without Aggregated

data (PISA) is proposed. The PISA model is adaptable to the distribution of the data and the location of the query boxes.

A performance study of four different index structures is performed in (17). The performance of index structures is evaluated with a set of nine parameters. Classification trees and an aggregation and scatter diagram method are the selected approaches for decision making.

4 Controlling with Uncertainty

Controllers build their decisions upon business figures. In most cases these business figures are treated as crisp data although measurement errors, estimation or other uncertainty aspects are common. Different approaches exists to include uncertainty into the decision process.

In (18) an expert system as a case study of a public transport system is evaluated.

In large data-sets exist logical, arithmetic probabilistic or more general structural relations between variables. If there are errors in the data then the underlying structural relations are violated. (LENZ91) focus on statistical quality control of data to detect contradictions between the data and the structural relations given a-priori.

Stochastic Controlling is presented in (19) and also in (20). Semantic consistency is research topic that is addressed in (21).

Uncertainty can also be expressed in Fuzzy sets. Work on Controlling with Fuzzy set theory is part of the work in (22) and (**?**).

Using multivariate distributions as the description of variables with errors and the usage of non-linear equation systems result in quite complex statistical environments. Markov Chain Monte Carlo techniques as a method for dealing with error-in-the-variables system are research in (23). An application of stochastic business figure systems with Balanced Scorecards is presented in (**?**).

That Fuzzy set approach and stochastic simulation are quite similar is shown in (KOEP06).

??? (25) ???

5 Decision Support

The decision making process is very complex and decision makers should be supported by techniques and software. A huge amount of methodologies exist. But which techniques delivers the most pragmatic solution is often not easy to evaluate.

Multi-criteria decision making or analysis (MCDA) is one of the techniques that were developed long time ago. But the growing markets of e- and m-commerce renewed the interests of these techniques. MCDA techniques are

reviewed in (26). A new hybrid techniques called "GiUnTa" is also presented there.

6 Data Mining

Along the mass of techniques, algorithms and methods that exist to work on data sets.

* Beitrag in Statistic Lexikon

(27)

7 Applications of Business Intelligence

Business Intelligence as a research area can only be justified if developed methods and techniques are tested and adapted to real world applications. In the following two out of many applications are selected.

7.1 Knowledge Management and Markets

Knowledge can be seen as an intangible asset that seems not tradable. The transfer of knowledge in electronic knowledge markets like virtual companies and strategic alliances is focus in (28). But the exhaustion of the potential is only reasonable if market compensations and quality assurance is considered.

Knowledge management inherits the major challenge to motivate people to share their knowledge. Companies use incentive systems to address this challenge. The combination of culture and incentive systems is analyzed in (29).

Mutual knowledge sharing can lead to a benefit for all the participants. However, establishing voluntary knowledge sharing can be difficult because each member benefits from the knowledge offered by others but gains little from the own contribution. The Data Trader Game (30) was designed and implemented for real-life experiments.

7.2 Travel Planning

In online traveling planning an active information system, which aims to notify a traveler timely about a likely delay is presented in (**?**). Besides providing the right content also the best notification time is included into the information system. The usage of inference diagrams is helpful to formulate and solve best choices for content and time.

8 Statistics

* Sampling * Quality Control
(31) (32) (33)
(34)

References

[1] Lenz, H.J.: M3-database design. manuskript. (1993) 1–36
[2] Lenz, H.J.: Zum entwurf statistischer datenbanken. Allgemeines Statistisches Archiv (1993) 60–67
[3] Lenz, H.J.: On the design of a statistical database, micro-, macro- and metadata modelling, historical social research. (1993) 1–28
[4] Lenz, H.J.: M3-database design - micro-, macro- and meta-database modelling. SoftStat '93. Advances in Statistical Software 5 (1994)
[5] Lenz, H.J.: A rigorous treatment of microdata, macrodata and metadata. Compstat 1994, Proceedings in Computational Statistics (1994)
[6] Lenz, H.J.: The conceptual schema and external schemata of metadatabases. Proceedings Seventh International Working Conference on Scientific and Statistical Database Management (1994) 160–165
[7] Lenz, H.J., Kssler, W., Boris, L.: On the distribution of read heads acting on a computer hard-disk. discussionpaper. (1993)
[8] Lenz, H.J., Shoshani, A.: Summarizability in OLAP and statistical data bases. In: Statistical and Scientific Database Management. (1997) 132–143
[9] Lenz, H.J., Thalheim, B.: Olap databases and aggregation functions. In: Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Washington, DC, USA, IEEE Computer Society (2001) 91–100
[10] Lenz, H.J., Thalheim, B.: Olap schemata for correct applications. In: TEAA. (2005) 99–113
[11] Neiling, M., Lenz, H.J.: Data integration by means of object identification in information systems. In: ECIS. (2000)
[12] Neiling, M., Lenz, H.J.: The german administrative record census an object identification problem. Allgemeines Statistisches Archiv **88** (2004) 259 – 277
[] Lenz, H.J., Köppen, V., Müller, R.M.: Edits - data cleansing at the data entry to assert semantic consistency of metric data. ssdbm **0** (2006) 235–240
[13] Michalarias, I., Lenz, H.J.: Dissemination of multidimensional data using broadcast clusters. In: ICDCIT. (2005) 573–584
[14] Michalarias, I., Boucharas, V., Lenz, H.J.: Hybrid scheduling for aggregated data delivery in wireless networks. In: Proceedings of the First

International Conference on Communications and Networking in China, Beijing, China (2006)

[15] Jürgens, M., Lenz, H.J.: The r_a^*-tree: An improved r-tree with materialized data for supporting range queries on olap-data. In: DEXA Workshop. (1998) 186–191

[16] Jurgens, M., Lenz, H.J.: Pisa: Performance models for index structures with and without aggregated data. In: 11th International Conference on Scientific and Statistical Database Management, Los Alamitos, CA, USA, IEEE Computer Society (1999) 78

[17] Jürgens, M., Lenz, H.J.: Tree based indexes versus bitmap indexes: A performance study. Int. J. Cooperative Inf. Syst. **10**(3) (2001) 355–376

[18] Lenz, H.J., Leuthardt, H.: Exbus - ein expertensystem zur betriebswirtschaftlichen analyse von verkehrsbetrieben. Entscheidungsunterstzende Systeme in Unternehmen (1988)

[19] Lenz, H.J., Mae, U.: Stochastisches controlling. Wissenschaftliche Zeitschrift TH Ilmenau Jg. 37 (1991) 157–167

[20] Lenz, H.J.: Controlling unter unsicherheit. Rechnungswesen und EDV : kritische Erfolgsfaktoren im Rechnungswesen und Controlling (1991) 359–364

[21] Lenz, H.J.: Semantic consistency of fuzzy data. First IASC World Conference on Computational Statistics and Data Analysis (1987) 59–65

[22] Lenz, H.J., Müller, R.M.: On the solution of fuzzy equation systems. In Riccia, G.D., Lenz, H.J., Kruse, R., eds.: Computational Intelligence in Data Mining. CISM Courses and Lectures. Springer, New York (2000)

[23] Köppen, V., Lenz, H.J.: Simulation of non-linear stochastic equation systems. In S.M. Ermakov, V.B. Melas, A.P., ed.: Proceeding of the Fifth Workshop on Simulation, St. Petersburg, Russia, NII Chemistry Saint Petersburg University Publishers (July 2005) 373–378

[KOEP06] Köppen, V., Lenz, H.J.: A comparison between probabilistic and possibilistic models for data validation. In Rizzi A, V.M., ed.: Proceeding in Computational Statistics COMPSTAT 2006, Rome (2006) 1533–1541

[25] Lenz, H.J.: Knowledge-based economic analysis of a public transport company. Diskussionsarbeit 3/1987 (1987)

[26] Lenz, H.J., Ablovatski, A.: MCDA - Multi-Criteria Decision Making in e-commerce. In: Decision Theory and Multi-Agent Planning. Springer (2006) 31–48

[27] Lenz, H.J.: On the idiot vs proper bayes approach in clinical diagnostic systems. technischer report. (1992) 1–8

[28] Müller, R.M., Spiliopoulou, M., Lenz, H.J.: Expertenrat in E-Marketplaces. In: Elektronische Marktplätze. dpunkt.verlag (2002) 38–48

[29] Müller, R.M., Spiliopoulou, M., Lenz, H.J.: The influence of incentives and culture on knowledge sharing. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), Los Alamitos, CA, USA, IEEE Computer Society (2005) 247b

[30] Müller, R.M., Haiduk, S., Heertsch, N., Lenz, H.J., Spiliopoulou, M.: Experimental investigation of the effects of different market mechanisms for electronic knowledge markets. In: ECIS. (2005)
[] Schaal, M., Lenz, H.J.: Best time and content for delay notification. In: TIME. (2001) 75–80
[31] Lenz, H.J., Rödel, E.: Statistical quality control of data. Operations Research '91 (1992) 341 – 346
[32] Lenz, H.J., Wetherill, Barrie, G.: Frontiers in statistical quality control 4. Frontiers in Statistical Quality Control 4 (1992)
[33] Kössler, W., Lenz, B., Lenz, H.J.: ber die robustheit von liebmanresnikoff-variablenprfplne. Diskussionsbeitrge zur Statistik und Quantitativen konomik (1993) 1 – 44
[34] Kössler, W., Lenz, H.J.: On the non-robustness of ml sampling plans by variables. Vth International Workshop on Intelligent Statistical Quality Control - Proceedings (1994) 83–93