

Web Farming and Data Warehousing for Energy Tradefloors

Carsten Felden, Peter Chamoni

*Universität Duisburg - Essen, Standort Duisburg, Fakultät 3 Wirtschaftswissenschaft, Institut für Logistik und Informationsmanagement, Lehrstuhl für Wirtschaftsinformatik und Operations Research, Lotharstraße 65, 47048 Duisburg, Germany
{felden, chamoni}@uni-duisburg.de*

Abstract

The recent liberalisation of the German energy market forced the energy industry to develop and install new information systems to support agents on the energy trading floors in their analytical tasks. Besides classical approaches of building a data warehouse to give insight into the time series to understand market and pricing mechanisms it is crucial to provide a variety of external data from the web. Weather information as well as political news or market rumors are relevant to give the right interpretation to the variables of a volatile energy market. Starting from a multidimensional data model and a collection of buy and sell transactions a data warehouse is built that gives analytical support to the agents. Following the idea of web farming we harvest the web, match the external information sources after a filtering and evaluation process to the data warehouse objects and present this qualified information on a user interface where market values are correlated with those external sources over the time axis.

1. Introduction

This paper presents a new approach to enhance data warehouses by adequate and highly related information from internet sources. An integrated process of searching and semi-automated evaluation will be developed to bring more precise information from outside into a data warehouse. The largest German utility company (RWE AG, Essen, Germany) runs a market information system to give highest possible analytical support to their energy traders. The kernel is a data warehouse which stores harmonized heterogeneous data derived from internal trading systems and external information brokers. The conceptual framework and the data model of this *Systematic Analysis and Research Tool* (SMART) for energy trading will be introduced in chapter 2. Until now there are only few published approaches which tackle this

problem domain and give approved solutions for the coupling of internal and external data. The classical approach is published by Hackathorn [6]. He suggests that external data must be qualified and classified by a person who works as an information editor within the company. A second approach derives from the idea of *Business Information Collection* as part of the strategic enterprise management (SAP SEM™) initiative [12]. This process of integrating external (business) information into the data warehouse isn't automated either but needs an editor workbench. In chapter 3 we will show three methods which enhance the idea of web farming. Firstly we build a set of meta data based descriptors to classify external information, secondly we train a filter (artificial neural network) to select potential interesting information and thirdly we implement a graphical user interface which connects the information sources via a *star field* to the time series stored in the data warehouse. Chapter 4 will summarize the findings and gives ideas for further developments.

2. Conceptual framework

The collection, reduction and selection of relevant information can only occur on the basis of consistent company-wide data retention. Due to the heterogeneous legacy systems a systematic bringing together of relevant databases is necessary. The data warehouse concept is an attempt to efficiently manage and collect relevant information derived from the vast amount of data.

2.1 The Data Warehouse Concept

First of all a formal definition of the term data warehouse seems appropriate: "... a data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of management's decisions." [7] The structure of a data warehouse is totally different from the structure of operational data bases. A data warehouse differs due to the objective of an operational data base by the type of the entered data and their supply.

The core of a data warehouse is a data base, in which data from different operational systems are historically saved in different levels of aggregation.

Due to the fact that, as a rule, analysts make complex inquiries and demand intuitive working with the database, a multidimensional data model seems appropriate. The complexity of a multidimensional structure is the result of the amount and the type of dimensions. Dimensions can be seen as the highest reduction level of data [5]. Therefore, two types of dimensions can be differentiated. On the one hand, all elements of a dimension are equal; this means they all have the same granularity. On the other hand, there is a hierarchical relationship between them [3].

The dimensions of the described multidimensional data model are the basis of the integration process of external information. If the retrieval query refers to dimension terms, the result must be linked to these dimensions. If the retrieval query results from individual dimension attributes, a coupling must be ensured to the explicit OLAP-slice, in order to guarantee the context of the inserted information.

2.2 Modelling the Data Warehouse

Acknowledging the necessity of a data warehouse which satisfies the information need of a company, the following chapter will present the market information system SMART of RWE Trading GmbH. For the energy tradefloor a market information system was developed with information from internal and external company sources. For the illustration of the results of the information analysis the *Application Design for Analytical Processing Technologies* (ADAPT) was selected as modelling method. In the context of the scope for discretion, which ADAPT offers, we only use as less different symbols as possible [3]. For pragmatic reasons one has to assume that the information was directly inserted into hypercubes. The only hypercubes which is presented here is the following: *Product Price History*. In this hypercube the dimensions *Product*, *Product_Type*, *Region*, *Commodity*, *Information_Source*, *Currency*, *Validity_Time*, *Transaction_Time* are combined. Figure 1 illustrates the ADAPT-model.

Aim of this analysis is the presentation of the respective bid-, ask- (including their average) and fixing-rate of the traders of electricity in the course of time. The rates can refer to different regions (e.g. CEPI {North Germany}, EIS {South Germany}) or to different types of products (e.g. base or peak characteristics). Valid times and transaction times are stored to document the actions and decisions of an agent.

In order to create the multi dimensional data cubes for all reports, 14 different dimensions were identified. These data cubes are the origin of fact tables and are modelled in

a Galaxy with 1:n-relationships between the dimension and fact tables.

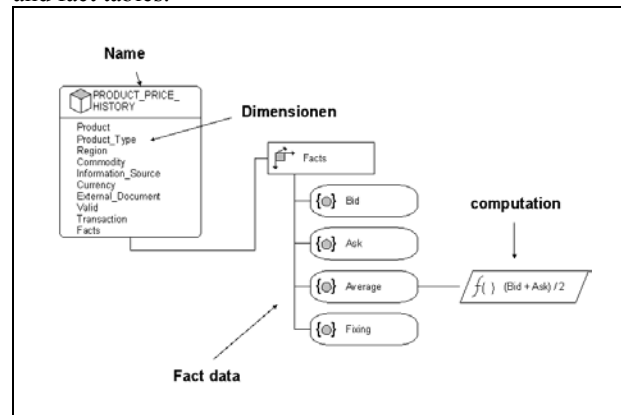


Figure 1: Product Price History

The implementation of the data model described above seizes data from internal as well as from external sources. External sources are data streams of external information suppliers (German Weather Services, Prebon or Intergrid), with which an appropriate contract was signed. These data streams represent time series of stock exchange rates or production and market data. Likewise power station usage data are stored as well as power station losses. Meteorological data allow predictions of energy consumption of customer groups in certain regions.

2.3 Information retrieval process

The goal of further research activities was to find appropriate sources in the internet and to make these sources available for SMART. The fully automated retrieval process is illustrated in the following figure 2.

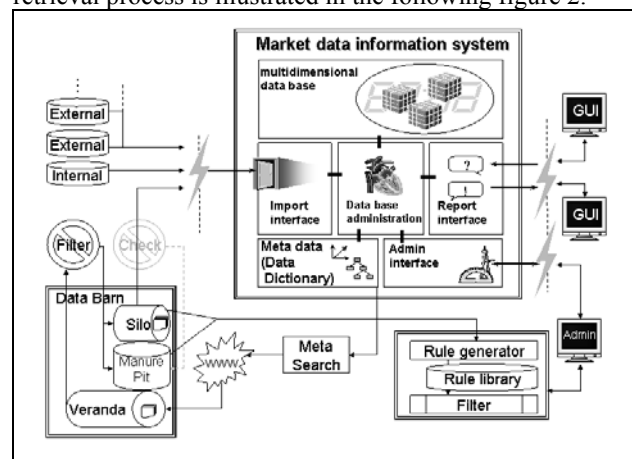


Figure 2: Retrieval process

With the aid of meta data from the data warehouse a Web extracting and loading tool is applied, in order to find and transfer information. Empirical studies have shown that the number of new internet pages is not increasing

dramatically. Therefore it is sufficient to start this process weekly. All information are stored in the so-called data barn. The data barn makes transient and persistent storage for the processing of internet pages available. A central problem for the integration is the extraction of information from the internet and the generation of rules for the classification of data. A Multilayer Perceptron (MLP) with data descriptors and data quality parameter as input is used. The process has shown that a maximum of 80 pages is identified as a set of results of the internet search. Just three pages are remaining after the application of the MLP with a classification correctness of 80 percent. Interesting internet pages are linked to the respective hypercubes to accomplish the integration process in order to make an overall analytical access possible [2][8][10][11]. This happens in form of a further dimension in the employed Star Scheme. The uninteresting internet pages remain as training data in the persistent 'manure pit'. Independent of manual examinations within the process, the evaluation of individual sites is almost in real-time. The evaluation time of an average text of 1,000 words is less than 20 seconds.

3. Method

Internet documents have to be identified, classified and evaluated to clarify their grade of interest [14]. As the identification of information is done by a meta search engine, the following part concentrates on information evaluation. In principle the evaluation refers to text documents, since an automatic contentwise evaluation of picture, audio and video files is not feasible yet without difficulties. For this purpose the construction of a *Rules Generator* is an essential problem. Rules have to be generated, which identify those internet data for the market information system in a filter exclusively, which are also actually relevant for decision makers.

3.1 Descriptor definition

In order to select a classification structure (information surrogate) both an operationalization of the quality of information and methods of text mining must be taken into consideration. The final aim is to identify actual interesting pages and offer these to decision makers. The information surrogate consists of a vector capturing a subset of meta data, quality demands and conceptual extension. It is assumed that text information from the internet is already available in a company and will be divided manually into interesting and uninteresting pages. The already existing pages are the training basis for identifying the descriptors that characterize interesting and uninteresting texts. In order to determine the importance of the descriptors their frequency is calculated [17][18][9].

Identical facts are presented differently on WWW-pages and accordingly the quality of information varies [16]. In order to determine the source type, the author of the website is taken into consideration. Therefore the type of organisation and financing (e.g. advertising income) must be analysed. The quality of the source is evaluated by correctness, actuality and navigation path.

In favour of the retrieval process it is useful to complete the identified descriptors of texts with terms which improve their classification. The underlying problem is that qualified pages can be classified as interesting by means of the determined descriptors even though they are not really meaningful to the decision maker. This means that further characteristics have to be added manually.

3.2 Filtering process

The filtering process is described by generated rules that are stored in a library. It is the aim to select the interesting information and then put it at the user's disposal. With help of a neuronal net a classification can be carried out. Therefore, the data is first entered into an input layer and then further processed. The input values must be coded for the neural network accordingly, so that the rules can be applied. The output is a classification indicating whether an internet page is interesting or not. The classification is done by a Multilayer Perceptron (MLP). This MLP consists of an input layer, one hidden layer and an output layer. A perceptron takes over the input from the preceding unit. Input values are the descriptors, their completions as well as the quality characteristics. Thereby w_{ji} is the weight of the value in the input unit for the transition to the hidden unit j . The net input net_{2j} is determined as the sum of the weighted inputs. The activation function $a_{2j}(net_{2j})$ computes the activation. The result of the application of the neuronal net is finally a classification of internet pages whether they are either interesting or uninteresting. The multiplier of the output function is in $\mathfrak{R}[0, 1]$ (considering a threshold value). The data search and data evaluation, done by the filter, can be understood as a stationary information agent [13][19].

3.3 Visualization

A central problem consists in putting the data from heterogeneous sources into a single ultimately understandable user interface. Visualization needed in the market information system differs from the conventional concepts of generating inquiries and information retrieval by the ability of fast filtering, progressive change of search parameters and continuous change of goals and visual analysis of identified results. This leads to the

following points, which must be implemented in the market information system: combining dynamic query filter, star field display and tight coupling [1][14][15]. The display is illustrated in figure 3.

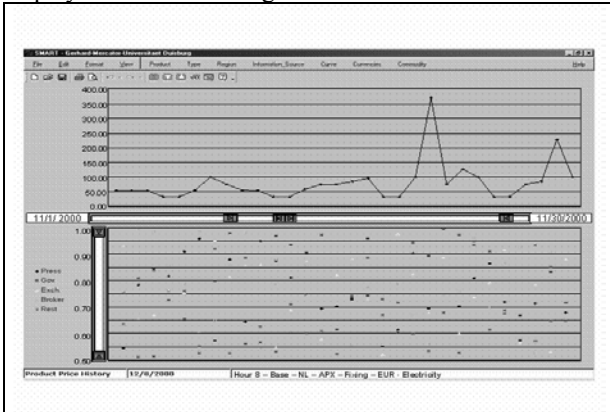


Figure 3: User interface with the star field display

The screen is splitted by a time slider in the center. The upper diagram represents the market data from well-known sources in a curve shape diagram. The lower diagram represents the stored internet documents in form of a star field display (the y-axis reflects the measure of interest). Both diagrams are adjusted to the validity time of the data via time slider. The validity time of the internet documents results from the date, at which the document was placed on the respective internet server. The status bar on the bottom shows the selected dimension values. It is possible to change the upper and lower frame representations to full size for further analysis. User studies have shown that early user integration by scenario and storyboard technique supported a broad acceptance and an intuitive usage of the system.

4 Conclusion

The development of SMART shows that the integration of internal and external data is crucial. However, first of all energy companies have to implement analytic information systems within their enterprises to gain benefit of this architecture. The usage of the market information system shows that the database improves the analytical power of decision makers, in order to recognize tendencies in the energy market promptly. Nevertheless the respective model and the system must grant a high flexibility to adjust them to changing conditions in the energy market. Furthermore the activities on the energy market and the work of the analysts will enhance the system. Market information systems have to be optimized by better evaluation of external information and automatization of process integration.

References

- [1] Ahlberg, C.; Shneiderman, B., "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays", in: Card, S. K., J. D. Mackinlay, and B. Shneiderman, (ed.), *Readings in Information Visualization - Using Vision to Think*, San Francisco, 1999, pp. 244 – 252.
- [2] Bishop, C. M.: *Neural Networks for Pattern Recognition*, Oxford, 1995.
- [3] Bulos, D., "OLAP Database Design - A New Dimension", in: Chamoni, P., P. Gluchowski, (ed.), *Analytische Informationssysteme - Data Warehouse, On-Line Analytical Processing, Data Mining*, Berlin, 1998, pp. 251 - 261.
- [4] Card, S. K., J. D. Mackinlay, and B. Shneiderman (ed.), *Readings in Information Visualization - Using Vision to Think*, San Francisco, 1999.
- [5] Codd, E. F., *OLAP On-Line Analytical Processing mit TM/1*, Whitepaper, 1994.
- [6] Hackathorn, R. D., *Web Farming for the Data Warehouse*, San Francisco, 1998.
- [7] Inmon, W.-H., *Building the Data Warehouse*, 3rd Edition, New York, 2002.
- [8] Khanna, T., *Foundations of Neural Networks*, Reading, Massachusetts, 1990.
- [9] Lalmas, M., Ruthven, I., "A Model for Structured Document Retrieval: Empirical Investigations", in: Fuhr, N., G. Dittrich, K. Tochtermann, (ed.), *Hypertext - Information Retrieval - Multimedia '97, Theorien, Modelle und Implementierungen integrierter elektronischer Informationssysteme*, Konstanz, 1997, pp. 53 - 66.
- [10] McCord Nelson, M.; W. T. Illingworth, *A practical guide to neural nets*, Reading, 1990.
- [11] McCulloch, W. S., Pitts, W., "A logical calculus of the ideas immanent in nervous activity", in: *Bulletin of mathematical biophysics*, Vol. 5, 1943, pp. 115 - 133.
- [12] Meier, M., Mertens, P., "The Editorial Workbench – Handling the Information Supply Chain of External Internet Data for Strategic Decision Support", in: *Journal of Decision Systems* 10/2, 2001, pp. 149 – 174.
- [13] Müller, J. P.; Wooldridge, M. J., and Jennings, N. R., *Intelligent Agents III - Agent Theories, Architectures, and Languages*, Proceedings of ECAI'96 Workshop (ATAL), Budapest, Hungary, August 12-13, 1996, Berlin, 1996.
- [14] Nayer, M., "Achieving Information Integrity: A Strategic Imperative", *Information Systems Management*, Vol. 10, No. 2, 1993, pp. 51 - 58.
- [15] Nielsen, J., *Usability Engineering*; Boston, 1993.
- [16] Redman, T. C., *Data Quality – the field guide*, Burlington, 2001.
- [17] Salton, G.: *Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer*, Reading, 1989.
- [18] Salton, G.; M.-J. McGill, *Introduction to Modern Information Retrieval*, Hamburg, 1983.
- [19] Wooldridge, M. J., J. P. Müller, M. Tambe, *Intelligent Agents II - Agent Theories, Architectures, and Languages*, IJCAI'95 Workshop (ATAL), Montréal, Canada, August 19 - 20, 1995, Proceedings, Berlin, 1995.