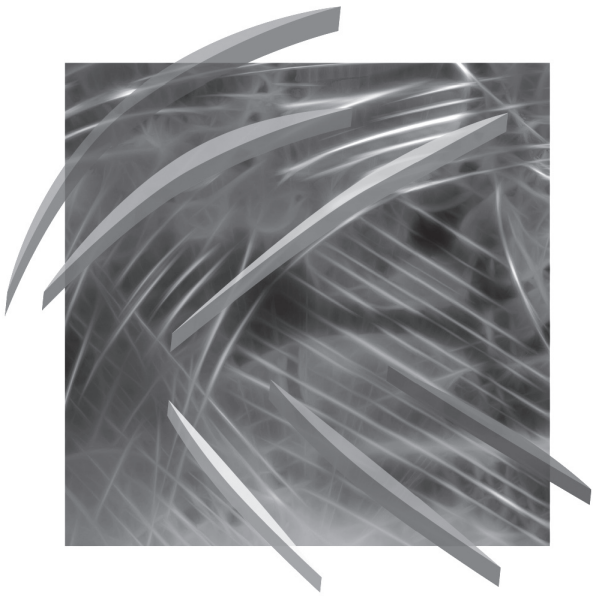
INFORMATYKA EKONOMICZNA
BUSINESS INFORMATICS

21 • 2011



Publishing House of Wrocław University of Economics
Wrocław 2011

Copy-editing: Agnieszka Flasińska, Elżbieta Macauley, Tim Macauley,

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com

Information of submitting and reviewing papers is available on the Publishing House's website www.wydawnictwo.ue.wroc.pl

ISSN 1507-3858 (Business Informatics)
ISSN 1899-3192 (Research Papers of Wrocław University of Economics)

The original version: printed

Printing: Printing House TOTEM
Print run: 200 copies

Contents

Streszczenia

INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 21 • 2011

ISSN 1507-3858

Paweł Weichbroth
Memex

LOGICAL DATABASE DESIGN FOR MARKET BASKET ANALYSIS

Abstract: Data is a key component of developing an Expert System for market basket analysis. Getting data safely stored, analyzed and updated in databases is a crucial factor for discovering knowledge and its quality. This article presents a logical data model which is going to be implemented in a physical environment. We started off by gathering the system's requirements from the user and presenting them with the use of UML notation. In connection with requirements, the Entity-Relationship (ER) Model for Market Basket Database was elaborated. Next, the model was evaluated in scope of user satisfaction. At the end of this paper, conclusions point out further research plans for developing the decentralized expert system for market basket analysis.

Key words: knowledge discovery, expert system, association rules, market basket analysis.

1. Introduction and motivation

Market basket analysis (MBA) known also as the affinity analysis is a next step in the development of the retail trade and its promotion. It shows customers' habits and gives information about combinations of purchased goods. Due to the analysis, it is possible to identify the goods which were bought together, when they were bought and in what amount. The advanced applications of the implementation operating in a real time enable current analysis. Association rules are used in order to describe relations between goods. Such an interactive analysis can be a decisive element in gaining a competitive advantage.

The aforementioned association rules can be applied in a few areas. One of them is the **sale analysis**. Below, two definitions are depicted which implicate **a posteriori** nature of the conducted study and as a result- the process of knowledge extraction, given in the form of **association rules**. A survey conducted in the USA in 2008 by the FactPoint Group [2008] may be the source of valuable knowledge for similar entities operating on the Polish market. It was carried out on the group of 50 largest retailers regarding the sale revenue from 400 million USD to 24 billion USD. The authors indicate in their report that the market basket analysis gives measurable effects in the form of:

– effective, more profitable promotional campaigns of goods and services,

– more precise products positioning on the market shelves,
– enhancing the quality of loyalty cards promotion,
– higher value and size of individual purchases,
– lowering the costs connected with supply chain management.

The database life cycle integrates the basic steps involved in designing a global schema of the logical database, allocating data across a computer network and ultimately defining local DBMS-specific schemas. Taking the above into consideration, logical design is a second step in the entire process. Among the variety of data modeling approaches, we use the entity-relationship (ER) and UML data models. They are arguably the most popular ones in use today, due to their simplicity and readability [Teorey, Lightstone, Nadeau 2006].

In the article by P. Weichbroth [2010a] the expert system framework for market basket analysis was presented. Additionally, the overall functionality was presented. It is worth mentioning that the main system's components are: inference engine, knowledge database, market basket database. For the accurate physical implementation of database, the logical data model was designed and verified. We assume that this project is an important matter to undertake because of the complexity of the analyzed problem. A design approach is based on the studies of users' requirements in order to fulfill a suitable solution. It can be implemented in e-commerce solution in context of customer relationship management [Kubiak, Weichbroth 2010].

2. Definitions

The degree of interdependence between certain products can be presented in various ways. In the scope of descriptive statistics the standard measure is Pearson's Correlation coefficient. Taking the market basket analysis into consideration, it is restricted by two imposed variables (a, b) which are described in the study. On the other hand, the products that the market basket consists of have a binary distribution. That is why its application is impossible.

The standard association rules show the relation between certain goods. This relation consists of two items out of two sets of values: **predecessor** (*conditioning* or *premise*) – $\boldsymbol{A}$ and **consequent** (*conditional* or *conclusion*) – $\boldsymbol{B}$, which can be written as $\boldsymbol{A} \rightarrow \boldsymbol{B}$ [Cunningham, Frank 1999]. The rule is countable in the form of two measures, which show the uncertainty rate of the rule.

The first measure is the support ratio (*support*), which is the quotient of the number of transactions that include simultaneously the predecessor and the consequent to the number of all the transactions, the formula is as follows [Weichbroth, Korczak 2006]:

$$\text{support ratio}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N}. \tag{1}$$

The second measure is the confidence ratio (*confidence*), which is the proportion of the number of transactions including all the predecessor and consequent's transactions to the number of transactions including predecessor. The formula is as follows [Weichbroth, Korczak 2006]:

$$\text{confidence}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N_A} = \frac{\frac{N_{A \rightarrow B}}{N}}{\frac{N_A}{N}} = \frac{\text{support}\{A \rightarrow B\}}{\text{support}\{A\}}. \quad (2)$$

For example, for a very little data base sector including 100 transactions, in 10 of which the product *A* was bought and at the same time 5 of them, include product *B*, the association rule in the form $R(A \rightarrow B)$ has the support ratio 10% (10/100) and confidence ratio 50% (5/10). The low level of support (for example 1 transaction out of 100 thousand) shows that the certain rule is irrelevant or that the data include errors. The support ratio can be defined as a probability that the transaction chosen randomly from the data base will contain the predecessor and the consequent. The confidence ratio can be defined as a **conditional probability** that the transaction chosen randomly will contain all the predetermined elements of the consequent to the predecessor [Weichbroth 2010a, b; Weichbroth, Korczak 2006].

Association rules are output by algorithm in the following form:

```
<premise>→<conclusion><support value for rule>#<confidence
value for rule>
```

3. Requirements specification

Each solution has its particular strengths and weaknesses, which must be taken into account. In presented schema (Figure 1), for the purpose of the market basket analysis process, the database requires precisely defined attributes to keep track of sales activities. The requirements analysis for this database led to entities and their unique identifiers.

From the user's point of view, the system's requirements can be presented in the simple drawing as shown in Figure 1. On the other hand, the UML notation for use cases (Figure 2) involves stick figures representing a type of user and ovals representing each of the tasks that the user needs to be able to carry out [Churcher 2007].

Figure 1 can be perceived in the frame of knowledge discovery process from the sales transactions database. Each of the records in such a database commonly represents anonymous sales transaction with the specified quantity of each product. The simplistic view of these two records, in the user's perspective, was presented as two separate receipts (**Step 1. Data**). Each of the rows is inserted in the unique transaction table (stored in the database) by means of the … and the receipt printed

by register is given to a client. In the next stage (**Step 2. Information**), all frequent item sets are found. Based on them, association rules are generated (**Step 3. Discovered Knowledge**), which are structured representation form of knowledge. The implemented algorithm, widely described in [Mikulski, Weichbroth 2009], assumes that the minimum level of cut-off for support and/or confidence is given by the user. It has to be stressed that in case of the second measure, effective and reasonable implication of the cut-off takes place only for association rules. In the next stage (**Step 4. Knowledge Visualization**) discovered knowledge is visualized, using the conception of **rich link analysis**, presented by the author in [Weichbroth 2010a]. Ultimately, a user (expert) makes a decision (**Step 5. Knowledge Evaluation**) according to accept or reject the discovered knowledge, taking into account the domain problem and most importantly the purpose of the conducted analysis.

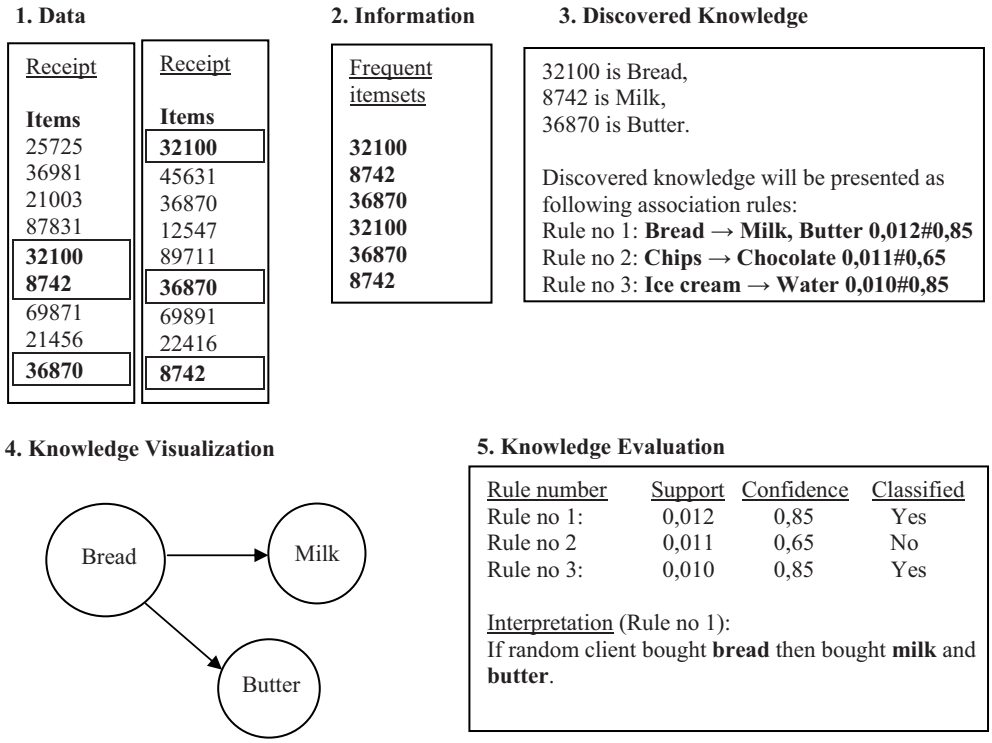


Figure 1. User's perspective: From raw data to evaluated knowledge

Source: own elaboration.

UML notation with use cases can be exploited to distinguish the principle functions of the required system (Figure 2). On the other hand, the notation itself is self-explanatory – it consists of simple drawings and plain text. It comprises an easy-to-do note of the interview between end-user and business analyst. The description

below gives a piece of information about what required functionality will be used to build a data model. Furthermore, based on that, input data can be specified for the system and on the other hand output data for the user.

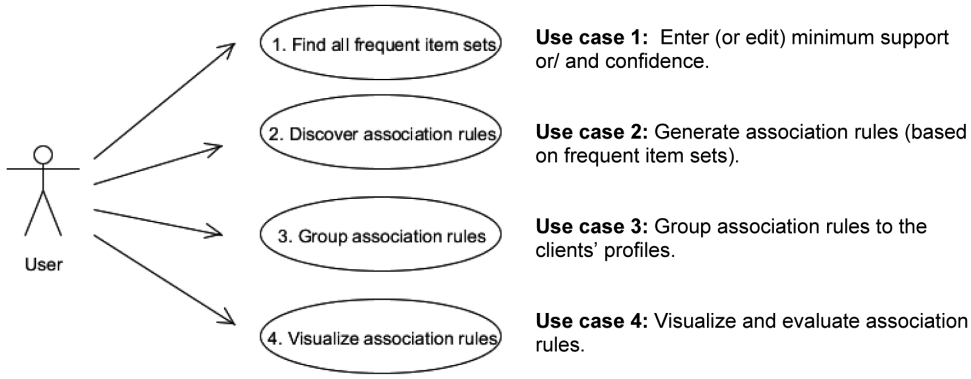


Figure 2. UML notation for use cases

Source: Own presentation[1]

4. Data model

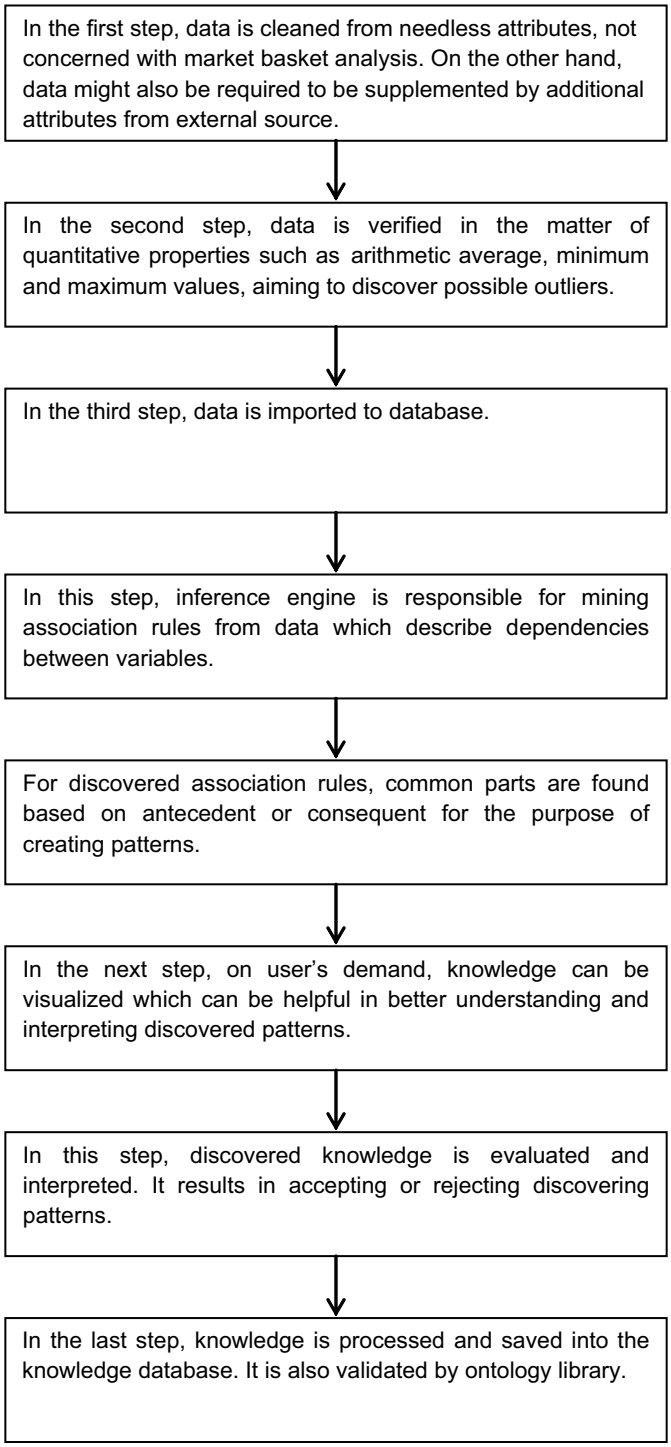
A practical way to start to get the idea for what the data involves is to draft an initial data model which is a representation of how the different types of data interact.

In general, there are two types of data models: a logical and a physical [Teorey, Lightstone, Nadeau 2006]. Presented in this work **logical data model** (LDM) is a complete representation of data requirements and the structural business rules that govern data quality in support of a project's requirements. The LDM shows the specific entities, attributes and relationships involved in a business function and is the basis for the creation of the physical data model. It is designed to hold information relating to a market basket analysis.

A practical way to start to get a feel for what the data involves is to divide the analysis into particular steps (Figure 3). We used a UML diagram which gives a clear view of sequences and conditions which must be fulfilled at each step. In parallel, every step has got a detailed explanation. Based on that, the unique functions stand out with respect of order against each other.

For this scenario we need to define the following facts (Fig. 4): *transaction*, *transaction item*, *vendor*, *shop*, *product*, *prices*, *warehouse*, *product group*, *frequent itemset*, *antecedent*, *consequent*, *association rules* and *pattern*. These facts define the requirements which the Database must meet and should be agreed between the Expert and the Database Designer prior to physical creation.

[1] The diagram in this article was prepared using UMLet 10.4 [http://www.umlet.com].


In the first step, data is cleaned from needless attributes, not concerned with market basket analysis. On the other hand, data might also be required to be supplemented by additional attributes from external source.
In the second step, data is verified in the matter of quantitative properties such as arithmetic average, minimum and maximum values, aiming to discover possible outliers.
In the third step, data is imported to database.
In this step, inference engine is responsible for mining association rules from data which describe dependencies between variables.
For discovered association rules, common parts are found based on antecedent or consequent for the purpose of creating patterns.
In the next step, on user's demand, knowledge can be visualized which can be helpful in better understanding and interpreting discovered patterns.
In this step, discovered knowledge is evaluated and interpreted. It results in accepting or rejecting discovering patterns.
In the last step, knowledge is processed and saved into the knowledge database. It is also validated by ontology library.

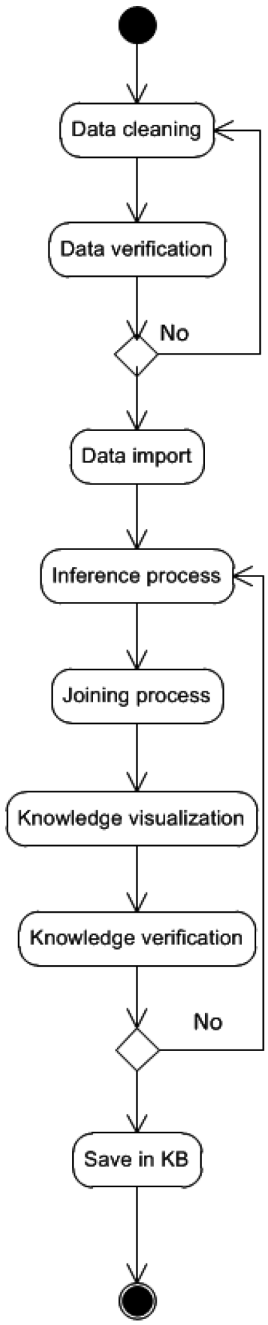


Figure 3. UML Diagram: Particular Steps in Market Basket Analysis

Source: own elaboration.

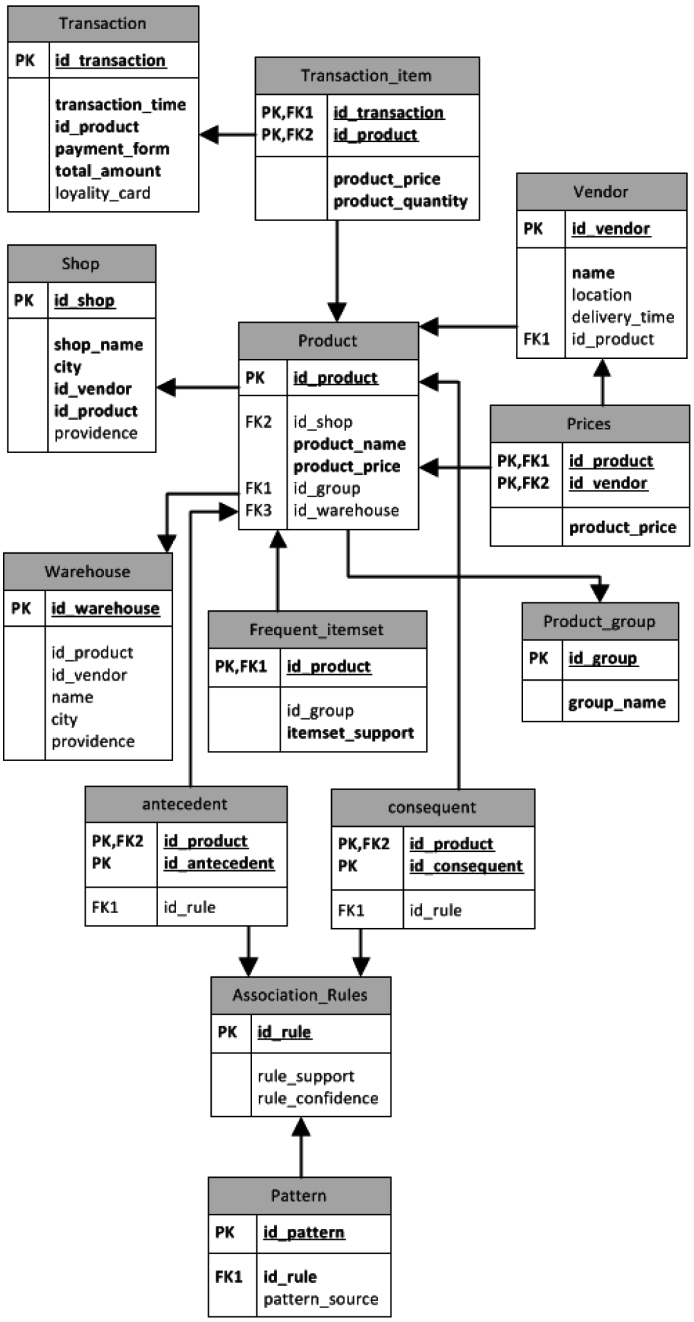


Figure 4. Entity-Relationship (ER) Model for Market Basket Database

Source: own elaboration.

The data-modelling diagram (Figures 3, 4) provides very precise and easy to interpret documentation. Even more importantly, the presented model is still easy to modify before physical implementation. This diagram allows the user to take in a large amount of information at a glance, giving him/her the ability to quickly get the idea of a database design without having to read a lot of text. We can conveniently satisfy the requirements of all the use cases (Fig. 2). Moreover, not all requirements were given by a user which are required by a developed expert system.

The first step in the model of market basket analysis is data preparation and its verification. If the data is correct then it can be processed and imported to the database. In the next step, the inference engine, having opened connection with database, analyzes the collected data. Direct objectives are finding frequent itemsets, association rules and linking them together to patterns. All actions taken by inference engine interact with the user for the purpose of a comprehensible dialogue between system and user (or expert). The analysis' results (time of work, number of frequent itemsets, the number of association rules divided according to the number of sets, support and confidence, the number of processed records from database) are saved into the knowledge base.

5. Model evaluation

In the next step, after the logical design of the model, we need to check whether the presented model is able to satisfy the requirements of the three use cases shown in Figure 2. Evaluation of the presented model will be conducted based on the physical implementation and obtained results. Another essential matter is data integrity which concerning relational databases means accuracy, correctness and validity.

6. Conclusions and future work

Dr Edgar F. Codd's landmark paper [Codd 1970], gave the computer science community a definition of a good database structure and a process for normalizing data by sequentially applying rules to the tables and entities. The presented model was verified due to the first three normalization rules. Obviously, there is no guarantee that later some attributes might be added to any table or even more – new tables to database.

Further research work will focus on the implementation of the agent system using Java Agent Development Framework (JADE).[2] The developing platform will support an enterprise in such areas like marketing, sales, logistics and price policy. So far, the author presented the algorithm for discovering association rules from database [Weichbroth 2009] and its modification [Mikulski, Weichbroth 2009]. The last work focused on the two methods which are being successfully used in

[2] More information is available on the website: http://jade.tilab.com/.

e.g. E-commerce activities [Kubiak, Weichbroth 2010]. Still, there is one issue to validate – the developing platform will work in an experimental environment which means that the system will be tested using random data. The real challenge is to deploy the platform in a real-time environment which can only objectively evaluate the described model and platform as well.

References

Churcher C., *Beginning Database Design. From Novice to Professional. Designing Databases for the Desktop and Beyond*, Apress, Springer-Verlag, New York 2007.

Codd E.F., A relational model of data for large shared data banks, *Communications of the ACM* 1970, Vol. 13, No. 6, pp. 377–387.

Cunningham S.J., Frank E., *Market Basket Analysis of library circulation data*, [in:] *Proceedings of 6th International Conference on Neural Information Processing, Perth, WA, Australia*, IEEE Computer Society, Perth 1999.

FactPoint Group. 2008, *Leading Practices in Market Basket Analysis. How Top Retailers Are Using Market Basket Analysis to Win Margin and Market Share*, http://factpoint.com/pdf2/1.pdf.

Kubiak B.F., Weichbroth P., Cross- and up-selling techniques in e-commerce activities, *eCommerce, ePayments and New Entrepreneurship* 2010, Vol. 15, No. 3, pp. 217–225.

Mikulski Ł., Weichbroth P., *Discovering patterns of visits on the Internet web sites in the perspective of associative models*, Polish Journal of Environmental Studies 2009, Vol. 18, No. 3B, pp. 267-271.

Teorey T., Lightstone S., Nadeau T., *Database Modelling and Design*, Elsevier, San Francisco 2006.

Weichbroth P., *Odkrywanie reguł asocjacyjnych z transakcyjnych baz danych*, [in:] A. Nowicki, I. Chomiak-Orsa (Eds.), *Rynek usług informatycznych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 82, Informatyka Ekonomiczna 14,Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław 2009, pp. 301–309.

Weichbroth P., *A framework of rule based expert system for market basket analysis*, [in:] J. Korczak, H. Dudycz, M. Dyczkowski (Eds.), *Advanced Information Technologies for Management – AITM 2010*, Research Papers of Wrocław University of Economics No. 147, Wrocław University of Economics, Wrocław 2010a, pp. 276-291.

Weichbroth P., *Zastosowanie reguł asocjacyjnych w analizie koszyka zakupów*, Marketing i Rynek 2010b, No. 9, pp. 27–29.

Weichbroth P., Korczak J., *Data mining – drążenie danych*, [in:] J. Korczak, M. Dyczkowski (Eds.), *Informatyka ekonomiczna. Część I. Propedeutyka informatyki. Technologie informacyjne*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 2006, pp. 243-268.

Websites

http://www.umlet.com.
http://jade.tilab.com/.

PROJEKT LOGICZNEJ BAZY DANYCH DO ANALIZY KOSZYKA ZAKUPÓW

Streszczenie: Dane są kluczowym komponentem systemów ekspertowych do analizy koszyka zakupów. Otrzymanie danych bezpiecznie zachowanych, analizowanych oraz aktualizowanych w bazach danych jest zasadniczym czynnikiem odkrywania bazy wiedzy i jakości tej bazy. W niniejszym artykule zaprezentowano logiczny model danych, który ma być zaimplementowany w środowisku fizycznym. Rozpoczynamy od zgromadzenia wymagań użytkownika i zaprezentowania ich w postaci diagramu UML. Wypracowany został model Entity-Relationship (ER) dla bazy danych o koszyku zakupów. Następnie model był oceniony pod względem zadowolenia użytkownika. W części końcowej artykułu przedstawiono wnioski i dalsze plany badawcze dotyczące rozwoju zdecentralizowanego systemu ekspertowego dla analizy koszyka zakupów.

Słowa kluczowe: odkrywanie wiedzy, systemy ekspertowe, reguły asocjacyjne, analiza koszyka zakupów.