PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

232

Knowledge Acquisition and Management

edited by
Małgorzata Nycz
Mieczysław Lech Owoc

Publishing House of Wrocław University of Economics
Wrocław 2011

Reviewers: Grzegorz Bartoszewicz, Witold Chmielarz, Halina Kwaśnicka,
Antoni Ligęza, Stanisław Stanek

Copy-editing: Marcin Orszulak

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Łopusiewicz

Typesetting: Beata Mazur

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 as well as in the annotated bibliography of economic issues BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

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

ISSN 1899-3192
ISBN 978-83-7695-200-0

The original version: printed

Printing: Printing House TOTEM

Contents

Presentations

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 232
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
Knowledge Acquisition and Management ISSN 1899-3192

Paweł Weichbroth
SynergiaIT

THE VISUALISATION OF ASSOCIATION RULES IN MARKET BASKET ANALYSIS AS A SUPPORTING METHOD IN CUSTOMER RELATIONSHIP MANAGEMENT SYSTEMS

Summary: An association rule in data mining is an implication of the form $A \rightarrow B$, where A is a set of antecedent items and B is a set of consequent items. In other words, it is defined on items which reflect the many-to-many relationships among items. Visualization has a quite long history of making large data sets more accessible using different techniques. Although many tools have been developed to visualize association rules, just few of them are able to filter by support and confidence ratio as well as key words. Moreover, few of these tools can handle a great number of rules with multiple antecedents. Thus, it might be difficult to visualize and understand discovered knowledge as a whole even when all rules are available. This paper presents the GraphMiner application to manage many of these problems. The application presentation is shown in the case study, where we discuss the opportunities on how it might be valuable in the CRM system due to support sales and promotion activities.

Keywords: knowledge visualisation, association rules, market basket analysis, Customer Relationship Management.

1. Introduction and motivation

Globalisation, market saturation, and growing competitiveness are the reasons for observed changes among relationships between a client and a supplier. The quality of offered goods and services is also considered in the context of maintained relations between two sides of a transaction. The concept of customer relationship management is the basis for creating the organisational structures, monitoring and supervising systems, implementing new products on the market, and promoting a trademark. Information given by the reports of various kind, the source of which is data processed in the information system, are of the vital importance in the strategic management. In the literature on the subject of management it is defined as [Penc 2001]: "management process intended for formulating and implementing strategies which promote better compatibility of anorganisation with its surrounding and accomplishing the strategic aims". In the model of strategic management system [Sopińska 2001], strategy designing was distinguished as the second stage and within variants assessment was extracted. Synthetic reports which are generated on the basis of the data obtained from organisation's operating performance can be a precious source for strategic variants assessment and its consequence for long-term aims.

The system of Customer Relationship Management is a general concept, which is difficult for unambiguous defining. Popularly, it is identified with dedicated computer system using the advanced IT technologies. It is the narrow insight – the extensive one is given by Swift [2001], who claims that CRM "is an iterative process that turns customer information into customer relationships through active use of and learning from the information". It gives the answer for changes taking place on the market such as lower level of customers loyalty, higher rotation of trade departments' employees, or higher cost of new clients acquisition. The concept of Customer Relationship Management implies [Stachowicz-Stanusch, Stanusch 2007] that what seems the most important is building the permanent relationships with customers. One of the aims of implementing CRM in a company is to achieve highly effective promotion (maximum return of promotion investment) as well as increase total sales to already existing customers.

Inferring from the survey done by Gemini Ernst&Young [Stachowicz-Stanusch, Stanusch 2007] taken on the group of 500 enterprises (300 from Eastern Europe, 200 from the USA), which employed at least 500 people, it can be stated that the highest rated aims (from 1-5, where 1 is the least important, while 5 means the most important) were the personalised offers for customers (4.4), increasing the loyalty of customers (4.4) and better customer understanding(4.3). The aforementioned aims can be realised by means of market basket analysis. To confirm the above, the survey of the FactPoint group must be quoted [Weichbroth 2010] – conducted in the US, in 2008. They took into account 50 largest retailers regarding the issue of sales revenue. The authors indicate in their report that market basket analysis allows achieving measurable effects in the form of [Weichbroth 2010]:

– effective, more profitable promotional campaigns as well as promotional goods and service;
– higher value and size of individual purchases;
– stating the "ideal" price point for individual products.

The degree of interdependence among defined products can be presented in different ways. Association rules have a few applications, among others, market basket analysis [Weichbroth 2010; Weichbroth, Korczak 2006]. Their graphic visualisation shows in an easy and accessible way found relationship between bought products. Such a form of presentation can be an effective tool offered within CRM, used in the work of call centres offices while doing promotion or selling goods or services. The aim of this work is to present the developed application which possesses the aforementioned functionality and indicate possibilities of its usage in the market basket analysis in the scope of Customer Relationship Management. This application is author's contribution, which will be the part of Multi-Agent System.

2. The basic definitions

In the scope of market basket analysis, the necessary terminology of frequent itemsets and association rules will be defined. Mining frequent itemsets and association rules were originally proposed in two papers by Agrawal, Imielinski, Swami [1993] and Agrawal, Srikant [1994]. A database D is a set of retail transactions which are sets over finite item domain. An itemset is a set of items, named more concisely. An itemset with k items is referred to as a k-itemset. The support (or support ratio) of an itemset A, denoted by sup(A), is the percentage $P(A)$ of retail transactions which contain it. However, we are just interested in mining frequent itemset which is an itemset whose support is above or equal to a minimum support specified by a user.

The association rule provides information in a form "if–then". They consist of two parts: antecedent, if (conditioning) A, and consequent, then (conditioned) B, which can be written in the form of $A \rightarrow B$ [Weichbroth 2009], where A and B are itemsets and $A \cap B = Ø$. The rule is countable in the form of two measures: *support* and *confidence* expressing the degree of *uncertainty* [Weichbroth 2010]. The *support* sup($A \rightarrow B$) of the association rule is the support of $A \cup B$, that is, P($A \cup B$). The *confidence* conf($A \rightarrow B$) of the association rule is the probability of B in condition of A, where conf ($A \rightarrow B$) = P($B \backslash A$) = P($A \cup B$)/P(A). The confidence value indicates how reliable a rule is. Therefore, the association rule is taken into account if its support and confidenceare larger or equal the minimum, both specified by a user, often defined as *cut-off*. This ensures a definitive result and it is one of the ways in which we can control the number of rules that will be generated. $P(B)$ is called *expected confidence* of the association rule $A \rightarrow B$. With the *lift* value which is defined as the factor by which the confidence exceeds the expected confidence, it is possible to interpret the importance of a rule. The lift value of a rule is defined as lift($A \rightarrow B$) = P($B \backslash A$)/P(B) or P($A \cup B$)/($P(A)$*$P(B)$) or simpler lift($A \rightarrow B$) = conf($A \rightarrow B$)/P(B). The expected confidence is equal to the support of the rule consequent. It is assumed in the definition of the expected confidence that there is no statistical relation between the rule antecedent and the rule consequent. This means that the occurrence of the antecedent does not influence the probability for the occurrence of the consequent and *vice versa*. In other words, the lift is a measure for the deviation of the rule from the model of statistic independency of the antecedent and consequent. Greater lift value (>>1) indicates stronger associations. In general, measures like support, confidence, and lift are also called *interest measures* because they facilitate with focusing on potentially more interesting rules [Hahsler, Chelluboina 2011].

Another important concept in mining frequent itemsets and association rules is *item group*. An item group is a closure of itemsets, that is, the union of all itemsets two or more of which share at least one item. Item groups are distinct – do not share items [Yang 2003].

3. Related work

The analysis of association rules is used in a variety of ways, i.e., banking, climate prediction, market basket analysis, merchandise stocking, insurance fraud investigation. This method is available in *R*, as the extension package *arules*. *R* is a language and environment for statistical computing (i.e., classification, classical statistical tests, clustering, linear and nonlinear modelling, time-series analysis) and graphics [Teetor 2011]. In their work, Hahsler and Chelluboina [2011] presented a few examples (see Section 4) based on a scatter plot where the variables on the axes are changed and consequences are discussed. This straightforward visualisation gives a high-quality view on the discovered knowledge. For exploration purpose, the scatter plot method offers interactive features for selecting and zooming.

Commercial data mining systems such as: *Statistica Data Miner* [Łapczyński 2009] from Statsoft, which is intended to be used in the whole process of data mining – starting from building queries to the database and finishing on the report; *SAS Visual Data Discovery* [SAS... 2011], where all graphics interact via the data table with brushing, highlighting, hiding, and other capabilities; Oracle Business Intelligence [*Oracle Business*... 2007] is a comprehensive business intelligence platform that delivers a full range of capabilities.

The matrix-based visualisation designs which position items on separate axes are among the most popular approaches to visualise binary relationships [Wong, Whitney, Thomas 1999; Yang 2003].

A directed graph is another prevailing technique to illustrate item associations. In our solution, the nodes of a directed graph represent the items, and the edges represent the associations. This technique is satisfactory enough when only a few nodes and edges are involved. An association graph can quickly turn into a confused mess with just a dozen rules [Klemettinen et al. 1994]. Hetzler et al. [1998] animate a directed graph to visualise disassociations and associations of information items. In two other works [Becker 1997; 1998], the author presented a series of smart visualisation techniques designed to support understanding of discovered knowledge. A comprehensive introduction of visualisation techniques provided by current data mining tools can be found in Westphal, Blaxton [1998]. In comparison to the aforementioned, the method in our solution enables flexible displacement of the whole association rules and their elements throughout the screen (see Figure 2). It assures tangible results, which means the greater clarity of association rules and their easier manipulation.

4. The market basket analysis

The issue of the market basket analysis was discussed in the works by Wick, Wagner [2006]; Han [2006]; and Weichbroth [2010], where the applications of association rules were presented. The binary analysis of market basket is data mining technique, whose aim is to find all those not known relationship between bought products. Adopted binary distribution [Wick, Wagner 2006] results from the fact that if the defined good was not bought, then the variable constitutes the 0 quality, and if it was bought (at any number), the variable constitutes the 1 quality.

The application of the market basket analysis can be found in the e-commerce solutions where typical cross-selling approach is used. Best examples might be the biggest on-line shops like amazon.com, bestbuy.com (the US), and merlin.pl (Poland). The dynamically generated hints, shown immediately after choosing a product, are the results of the analysis of the sales transactions' database, by finding all the sets of goods which were "frequent" bought together. The definition of frequency is arbitrarily precised by a user, defined as the support [Weichbroth 2010; Wick, Wagner 2006]. Mathematical definitions and detailed description of algorithm is discussed in the work by Mikulski, Weichbroth [2009].

5. The concept of CRM

The issues of CRM system were discussed in the works by Swift [2001]; Stachowicz-Stanusch, Stanusch [2007]. Stachowicz-Stanusch, Stanusch [2007] defined CRM as "set of strategies and methods whose main aim is to increase the loyalty of customers and decrease the costs of service, promotion and sales". The aforementioned quotation implies that CRM is a strategy; however, the IT system of CRM is only a tool which simplifies implementing the CRM. Without defining the strategy and stating the aims of CRM, the effect of its implementing can be only the consolidation of contractors' contact details, which in fact gives only an expanded telephone book. The concept of CRM states that what is the most important is not the profit maximising from an individual transaction but building permanent relationships with customers. There are a lot of reasons for such a purpose and they can relate to information need fulfillment, gathering knowledge, profit achieving, or cost reduction. The development of modern IT and telecommunication technologies enabled effective and efficient data processing. Basing on the above, CRM can be considered in two realms [Stachowicz-Stanusch, Stanusch 2007]: as overall activities in relation to acustomer and as an IT system supporting those activities.

Apart from the main purpose of building the loyal group of patrons, the CRM implementation, among others, aims to increase the effectiveness of marketing and promotion activities, which might lead to increase the volume of sales.

6. Application project

The GraphMiner applicationwasimplemented in ActionScript3 and the target platform for starting the application is any Web browser which supports Macromedia Flash. Data processing can be described in four steps:

Step 1: the script imports a data from the flat file.
Step 2: the script export a data to the database (MySQL).
Step 3: the script makes a XML files.
Step 4: the script generates charts which are based on XML files.

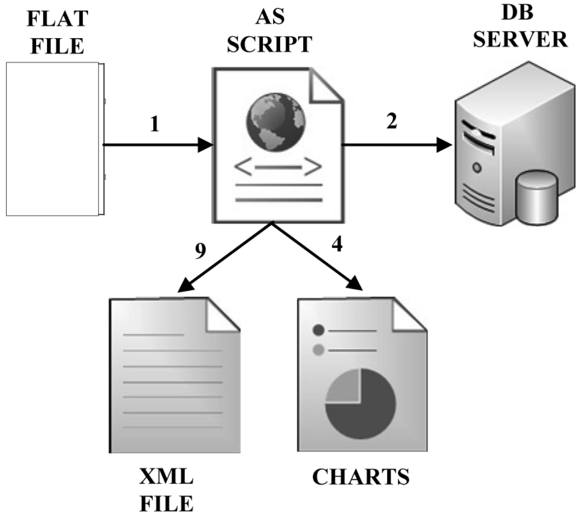

Figure 1. The process of generating the charts

Source: author's own study.

Input data for the importing script are the components of association rules, support, and confidence ratio. The schema of the programme was shown below in the form of three instruction manual:

1. The file containing the association relationship is loaded to the database by means of the importing file.

2. The generator of the XML file chooses data from the database, which recorded as association rules will be then visualised for a user. At this stage, the rules filtration takes place by stating the minimal value for support or confidence. Data filtration, based on the key words, from which the association rules are to be submitted, is also possible.

3. The script generating the chart induces automatically the generator of the XML file in order to take the data needed for display.

7. Charts generator

Data from the XML file are materialised in the script as two objects of two classes. The *Element* class stores the information about individual elements of association rules (the *antecedent* and *consequent*). Each object of the this class is graphically presented as a *dot*. The *Line* class is an object combining two classes of the *Element* kind (visually is the pointer combining the *antecedent* and *consequent*) representing association rules, assuming as parameters accordingly the support and confidence ratio for the thickness and colour of the line. The higher the support, the thicker the line is, and the darker the colour the higher the confidence. Parsing the XML file means:

– loading the *antecedent* and *consequent* and inserting to the table with the objects of *Element* class;
– loading the relationship between *antecedent* and *consequent* and inserting to the table with the objects of *Line* class;
– displaying all the objects of *Element* class;
– displaying all the objects of *Line* class.

8. Case Study

In order to verify this work's hypothesis, the case study will be presented in the scenario given below. For a small part of database, including 100 transactions, in ten of which a DVD player and in eight a TV set were bought and, at the same time, five of them included DVD and a TV set. Thus, the association rule of theform of *R(DVD Player→Tv)* has the **support** ratio **0.05** (5/100), **confidence** ratio **0.50** (0.05/0.1), and **lift 6.25** (0.5/0.08). The lift is a value between 0 and infinity and when it is greater than 1, this indicates that the antecedent and the consequent appear more often together than expected. This means that the occurrence of the antecedent has a positive effect on the occurrence of the consequent [IBM 2011].

The low level of support (for example 1 transaction out of 100 thousand) shows that the defined rule is irrelevant or that errors occur in the data. The support ratio can be defined as probability of occurring random transaction containing both the antecedent and consequent. Confidence ratio can be defined as probability of occurring random transaction containing all the predetermined elements of consequent to the antecedent [Weichbroth, Korczak 2006].

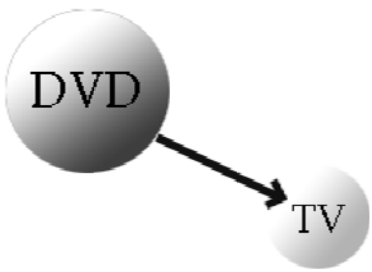

Figure 2. Visualisation of chosen association rules

Source: author's own study.

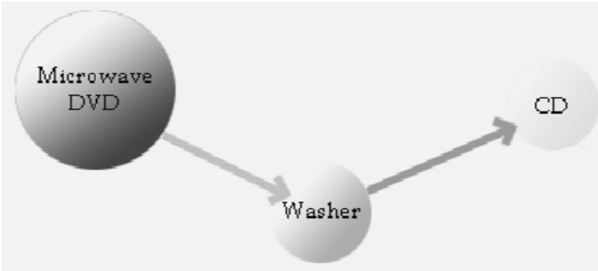

Figure 3. Visualisation of connected association rules

Source: author's own study.

The call centre office operating on the household goods market runs the telephone promotional campaign to which the buyers of DVDs were chosen. To answer the question which other goods they bought, a market basket analysis was performed. Based on the database sales, the transactions relationships between individual goods were found. Some of them were presented graphically in Figures 2 and 3. Taking into account prevalent association rule in the form of *R(Dvd Player→CD TV Microwave)*, the user of CRM and simultaneously the call centre's employee is able to obtain information about potentially interesting goods for the interlocutor. That gives higher chances for accepting an offer and complete a transaction.

Another view of discovered association rules is presented in Figure 3.

9. Conclusions and future work

In this paper, the application to visualise association rules was presented. So far, it has been broadly used in many data mining projects. The visualisation of the association rules in GraphMiner application is a simple way of presenting the found relationships between items. Nevertheless, it is still a concept which requires verification in a customer service office in production environment. From the technical point of view, an important issue, and not easy to undertake, is the integration with existing CRM system. This might be a challenge to connect two different solutions even if we have access to a source code. It could turn out that the implemented functionality is not sufficient enough.

On the other hand, the association rules charts can be used in analytical reports or in presentations. The database architecture of the system, based on open-source solutions, provides many tools for data handling. The separation of the data source from customer's application gives the possibility for easier data changes and allows running many heterogeneous projects at the same time.

References

Agrawal R., Imielinski T., Swami A. (1993), Mining association rules between sets of items in large databases, [in:] *Proceedings of 1993 ACM-SIGMOD Conference on Management of Data*, pp. 207-216.

Agrawal R., Srikant R. (1994), Fast algorithms for mining association rules, [in:] *Proceedings of 20th International Conference Very Large Databases*, Santiago.

Becker B.G. (1997), Volume rendering for relational data, [in:] J. Dill, N. Gershon (eds.), *Proceedings of Information Visualization'97*, IEEE Computer Society Press, Phoenix, pp. 87-90.

Becker B.G. (1998), Visualizing decision table classifiers, [in:] G. Wills, J. Dill (eds.), *Proceedings of Information Visualization'98*, IEEE Computer Society Press, Research Triangle Park, North Carolina, pp. 102-105.

Hahsler M., Chelluboina S. (2011), *Visualizing Association Rules: Introduction to the R-extension Package arulesViz*, http://cran.r-project.org/web/packages/arulesViz/vignettes/ arulesViz.pdf (date of access: 30.09.2011).

Han J., Kamber M. (2006), *Data mining. Concepts and Techniques. Second Edition*, Morgan Kaufmann, San Francisco.

Hetzler B., Harris W.M., Havre S., Whitney P. (1998), Visualizing the full spectrum of document relationships, [in:] *Proceedings of 5th International Society Knowledge Organization (ISKO) Conference*, Lille.

IBM (2011), *DB2 Business Intelligence*, http://publib.boulder.ibm.com/infocenter/db2luw/ v8/index.jsp?topic=/com.ibm.im.model.doc/c_lift_in_an_association_rule.html (date of access: 03.10.2011).

Klemettinen M., Mannila H., Ronkainen P., Toivonen H., Verkamo I. (1994), Finding interesting rules from large sets of discovered association rules, [in:] *Proceedings of 3rd International Conference of Information and Knowledge Management (CIKM'94)*, ACM, Gaithersburg,Maryland, pp. 401-407.

Łapczyński M. (2009), *Analiza koszykowa i analiza sekwencji – wielki brat czuwa*, Materiały konferencyjne z Seminarium zastosowania Statystyki i Data mining, Statsoft, Warszawa.

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

Oracle Business Intelligence Blog (2011), http://oraclebi.blogspot.com/2007/02/using-market-basket-analysis-to-add.html (date of access: 30.09.2011).

Penc J. (2001), *Strategiczny system zarządzania. Holistyczne myślenie o przyszłości. Formułowanie misji i strategii*, Agencja Wydawnicza Placet, Warszawa.

SAS Visual Data Discovery (2011), http://www.sas.com/technologies/analytics/statistics/ datadiscovery/index.html#section=3 (date of access: 30.09.2011).

Sopińska A. (2001), Rola systemu informacyjnego w procesie zarządzania strategicznego, [in:] R. Borowiecki, M. Romanowska (eds.), *System informacji strategicznej. Wywiad gospodarczy a konkurencyjność przedsiębiorstwa*, Difin, Warszawa.

Stachowicz-Stanusch A., Stanusch M. (2007), *CRM. Przewodnik dla wdrażających*, Wydawnictwo Placet, Warszawa.

Swift R.S. (2001), *Accelerating Customer Relationships: Using CRM and Relationship Technologies*, Prentice Hall, New York.

Teetor P. (2011), *25 Recipes for Getting Started with R*, O'REILLY, Sebastopol.

Vaníčková P. (2010), *Market Basket Analysis User Manual*, Web Site: Department of Software and Computer Science Education, Faculty of Mathematics and Physics, Charles University in Prague, Prague.

Weichbroth P., Korczak J. (2006), Data mining – drążenie danych, [in:] J. Korczak, M. Dyczkowski (eds.), *Informatyka ekonomiczna. Cześć I. Propedeutyka informatyki. Technologie informacyjne*, Wrocław 2006.

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

Weichbroth P. (2010), Zastosowanie reguł asocjacyjnych w analizie koszyka zakupów, *Marketing i Rynek*, nr 9, pp. 27-29.

Westphal C., Blaxton T. (1998), *Data mining solutions – Methods and Tools for Solving Real-World Problems*, John Wiley and Sons, New York.

Wick M.R., Wagner P.J. (2006), Using market basket analysis to integrate and motivate topics in discrete structures, [in:] *Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education*, ACM, New York, pp. 323-327.

Wong P.C., Whitney P., Thomas J. (1999), Visualizing association rules for text mining,[in:] *INFOVIS'99 Proceedings of the 1999 IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington.

Yang L. (2003), Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates, [in:] *Proceedings of the 2003 International Conference on Computational Science and its Applications: Part I*, Springer-Verlag, Montreal, pp. 21-30.

WIZUALIZACJA REGUŁ ASOCJACYJNYCH W ANALIZIE KOSZYKOWEJ JAKO METODA WSPIERAJĄCA SYSTEMY KLASY CRM

Streszczenie: Reguła asocjacyjna w drążeniu danych to implikacja w postaci A→B, gdzie A to zbiór elementów poprzedzających a B jest zbiorem elementów następujących. Innymi słowy jest zdefiniowana na elementach odzwierciedlających związki typu wiele do wielu między elementami. Wizualizacja ma dość długą historię w poprawianiu dostępności dużych zestawów danych do analizy przy pomocy różnych technik. Chociaż zostało opracowanych wiele narzędzi mających na celu wizualizację reguł asocjacyjnych, tylko niewiele z nich jest w stanie filtrować po współczynnikach wsparcia i ufności tak jak po słowach kluczowych. Ponadto, tylko niektóre z tych narzędzi mogą obsłużyć dużą liczbę reguł z wieloma poprzednikami. Tak więc, może być trudno zwizualizować i zrozumieć odkrytą wiedzę jako całość, nawet jeśli wszystkie reguły są dostępne. W artykule przedstawiono aplikację GraphMiner, która rozwiązuje wiele z tych problemów. Prezentacja aplikacji jest przedstawiona jako studium przypadku, w którym omawiamy sposób, w jaki można w systemie CRM wspierać sprzedaż i działania promocyjne.

Słowa kluczowe: wizualizacja wiedzy, reguły asocjacyjne, analiza koszykowa, CRM (*Customer Relationship Management*).