OLAP Recommender: Supporting Navigation in OLAP Cubes Using Association Rule Mining

Bohuslav Koukal[1], David Chudán[2], Vojtěch Svátek[3]

Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics
University of Economics, Prague
nám. W. Churchilla 1938/4, 130 67 Praha 3

[1]bohuslav.koukal@yahoo.co.uk
[2]xchud01@vse.cz
[3]svatek@vse.cz

Abstract: The OLAP Recommender tool automates the multidimensional data exploration process and recommends potentially interesting views on the data to the user. It integrates two data analytics methods – OLAP visualisation and data mining, the latter being represented by GUHA association rule mining. Algorithms implemented in the tool include automated data discretization, setup of dimensions' commensurability, automatic design of the data mining task based on the data structure, and mapping between the mined association rules and the corresponding OLAP visualisation. The system was tested with real retail data and with EU structural funds data. The experiments indicate that complementary usage of association rule mining and OLAP analysis identifies relationships in the data with higher success rate than the isolated use of both techniques.

Keywords: OLAP navigation, GUHA association rules, OLAP Recommender, guided analytics

1 Introduction

Self-service Business Intelligence and visualisation tools, operating over multidimensional data cubes specifying the target numerical values, *measures*, using combinations of *dimensions*, are today widely used for reporting, creating dashboards and answering the users' questions about their data. However, manual browsing and visualising the data brings some obvious problems. The user usually has certain prior understanding of the problem area, which guides him/her into the portion of the data s/he considers interesting. If there are other interesting portions in the data that s/he does not know about, it is very hard or even impossible to discover them. Another problem is a lack of completeness – it is impossible for a human to manually identify all interesting relationships in large multidimensional data just by browsing the data visualisations.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 46-50.

OLAP Recommender copes with these problems by recommending interesting views on multidimensional data, thus revealing strong relationships (indicating both trends and abnormalities) identified in the data. Compared to other proposals to combine data mining and OLAP [1, 2, 3, 4], it is designed to find these relationships in more specific subsets of the data and offers the possibility of extending to full potential of GUHA association rules. In contrast to general concept of recommending systems, OLAP Recommender is not based on the users' behaviour but solely on the relationships found in the data.

2 Searching for strong relationships in the data

2.1 GUHA association rules

For identifying strong and potentially interesting relationships in multidimensional data, OLAP Recommender uses association rule mining, specifically, the GUHA 4-ft method, described by Rauch and Šimůnek [5]. GUHA association rules can be considered as an extension and generalization of the association rules, originally introduced by Agrawal et al. [6]. As an addition to the traditional association rules, GUHA offers three important extensions useful for mining multidimensional data. Firstly, GUHA offers seventeen different interest measures in comparison to two (confidence and support) defined for the original association rules. Then, GUHA offers conditional association rules, making it possible to mine the rules in a subset of the data, while still maintaining the interest measures in relation to the whole dataset. Last, but not least, GUHA offers different types of coefficients (attribute value restrictions in the hypotheses) – e.g. multiple elements subsets, sequences or cuts, while traditional rules can use only a single-element subset as a coefficient.

2.2 Mining multidimensional aggregate data

Association rules were originally introduced for transactional data (market basket analysis) [6]. Chudán [7] introduced a different approach for using the association rules for mining aggregate data and based on his work, Koukal defined additional steps needed for automatically setting up the association rule mining task for the OLAP data [8]:

1. Measures discretization: Based on our experiments we proposed to discretize the measures to equal frequency intervals. We can discretize quite finely (to 20 and more intervals, based on the data volume), taking advantage of sequence coefficients offered by GUHA method.
2. Interest measures: We use the *base* and *above average dependency* (as GUHA counterparts to the more common support and lift) interest measures because of their easy interpretability and understandability.

3. Dimensions commensurability: Which data subsets can be compared together and which should not be compared in association rule mining?[1] Our research on automatic setting of the commensurability levels is currently ongoing; thus, this setting is done by the user in the current version of OLAP Recommender.

2.3 Visualising the results

OLAP Recommender does not visualise the association rules themselves; rather it visualises the part of the original data defined by the association rules. We use 2D column charts for the visualisation, thus constraining the maximal association rule antecedent length to 2 (as dimensions on the x-axis and in the legend) and the consequent length to 1 (as the measure on the y-axis) with a condition of unlimited length (as the data filter – slice or dice).

3 OLAP Recommender tool

OLAP Recommender was implemented as a web application in the ASP.NET framework and is running at http://connect-dev.lmcloud.vse.cz/Recommender. The workflow consists of three steps:

1. The user uploads the data as a CSV table or as RDF data semantically described by the Data Cube Vocabulary,[2] and defines their structure (dimensions and measures).
2. The user runs an association rule mining task. In this step the user can set the interest measures values and commensurability levels to get the most appropriate results.
3. OLAP Recommender displays the mined association rules, serving as a link to the data visualisation.

4 Experiments

OLAP Recommender was tested with two different datasets, whose attributes are displayed in Table 1: a retail dataset, described in more detail in Chudán's thesis [7], and a dataset[3] about European structural and investment funds (ESIF), prepared in the OpenBudgets.eu project.[4] We ran 5 mining tasks for the retail dataset and 8 tasks for the ESIF dataset with different settings. The tasks returned 46 association rules for the former and 114 rules for latter. Almost 70 % of the results in the retail dataset and

[1] Considering sales in a hypermarket as an example, not setting the commensurability levels would lead to many useless rules, pointing out the obvious: e.g. high sales of pastry and low sales of electronics. After proper commensurability level setting, we can receive more valuable rules, e.g., which pastry type sales prevail among the pastry products.

[2] https://www.w3.org/TR/vocab-data-cube/

[3] Available at https://github.com/openbudgets/datasets/tree/master/ESIF/2014/dataset

[4] http://openbudgets.eu

40 % in the ESIF dataset identified the highest or the lowest columns of the chart visualisations (i.e., anomalies and outlying values in the data).

5 Conclusions

By experiments with two different real datasets we proved the usefulness of the association rule mining and OLAP analysis combination, as the OLAP Recommender for the most test cases found more relationships, more interesting relationships and relationships in more parts of the cube, compared to using only self-guided OLAP analysis or using both methods separately.

Tab. 1: Retail and ESIF datasets differences summary

Characteristic	Retail dataset	ESIF dataset
Row count	34360	7039
Row interpretation	Sales of one product in one day	Single funded project
Dimension count	8	3
Measure count	1	3
Hierarchy	Two dimension hierarchies (product and time) with depth 4 and 2	Flat structure
Time dimension	Yes	No
Domain	Retail	Public fiscal data
Data form	Single table in .csv file	RDF data

The research has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833, OpenBudgets.eu.

References

1. HAN, J. OLAP mining: *An integration of OLAP with data mining.* In the Proceedings of the 7th IFIP Conference on Data Semantics, Leysin, Switzerland, 1997, pp 1-9.
2. IMIELIŇSKI, T., KHACHIYAN, L. and A. ABDULGHANI. *Cubegrades: Generalizing Association Rules.* Data Mining and Knowledge Discovery, v. 6 n.3, 2002, pp 219-257.
3. KAMBER, M., HAN, J. and J. CHIANG. Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes. In: *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD 1997)*. Newport Beach, CA, USA, 1997, pp 207-210.
4. ZHU, H. *On-Line Analytical Mining of Association Rules*. Master thesis. Simon Fraser University, Burnaby, British Columbia, Canada 1998. Available at: http://www.cin.ufpe.br/~jtalr/Mestrado/Thesis/zhu98line.pdf
5. RAUCH, J. and M. ŠIMŮNEK. *Dobývání znalostí z databází, LISp-Miner a GUHA*. 1st ed. Praha: Oeconomica, nakladatelství VŠE, 2014. ISBN 9788024520339.
6. AGRAWAL, R., IMIELIŃSKI , T. and A. SWAMI. Mining association rules between sets of items in large databases. *Proc. SIGMOD '93*, 1993, pp 204-216.

7. CHUDÁN, D. *Association rule mining as a support for OLAP*. Dissertation thesis. University of Economics, Prague, Faculty of Informatics and Statistics, 2015.
8. KOUKAL, B. *OLAP Recommender: Supporting Navigation in OLAP Cubes Using Association Rule Mining*. Master thesis. University of Economics, Prague, 2017.