# GUHA method and Association Rules

Jan Rauch
University of Economics, Prague, Czech Republic

**Extended abstract**

Association rules were introduced in the early 1990's with a goal to better understand the purchase behavior of customers in supermarkets [1]. Transaction data recorded by point-of-sale systems is analysed. We assume there is a set $I = \{i_1, \ldots, i_n\}$ of possible items of goods and set $D = \{b_1, \ldots, b_m\}$ of market baskets; it is $b_i \subset I$ for $i = 1, \ldots, m$. An association rule is commonly understood as an expression of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. An example of association rule can be {butter, cheese} $\rightarrow$ {bread} expressing that customers who buy butter and cheese also often buy bread.

There are two important measures of interestingness of association rules. The *confidence* is defined as $conf(X \rightarrow Y) = \frac{\text{number of baskets containing } X \cup Y}{\text{number of baskets containing } X}$ and the *support* is defined as $supp(X \rightarrow Y) = \frac{\text{number of baskets containing } X \cup Y}{m}$. A task of mining association rules is understood as a task of finding all association rules $X \rightarrow Y$ satisfying $conf(X \rightarrow Y) \geq minC$ and $supp(X \rightarrow Y) \geq minS$ in a given set of market baskets $D$. Here $minC$ and $minS$ are user-specified minimum confidence and support. This task is usually solved by the apriori algorithm [1] which has been many times implemented and modified.

The idea of association rules has been later generalized to data in a tabular, attribute-value form. The association rule is understood as an expression $Ant \rightarrow Con$ where $Ant$ and $Con$ are conjunctions of attribute-value pairs. Additional measures of interestingness of association rules have been defined [2].

However, the concept of association rules was introduced and studied already in 1960s in the framework of development of the GUHA method [3]. Monograph [4] introduces a general theory of mechanized hypothesis formation based on mathematical logic and statistics. Association rules introduced and studied in [4] are general relations $\varphi \approx \psi$ between general Boolean attributes $\varphi$ and $\psi$ derived from columns of an analysed data matrix. The symbol $\approx$ corresponds to a condition concerning contingency table of $\varphi$ and $\psi$.

If $A$ is a column and $\alpha$ is a subset of its possible values, then $A(\alpha)$ is a basic Boolean attribute. $A(\alpha)$ is true in a row $o$ of a data matrix if a value $A(o)$ of $A$ for the row $o$ belongs to $\alpha$. If $\varphi$ and $\psi$ are Boolean attributes, then $\neg\varphi$, $\varphi \wedge \psi$, and $\varphi \vee \psi$ are Boolean attributes. Their values are defined in a usual way. The term *association rules* has been used for relations $\varphi \approx \psi$ of general Boolean attributes $\varphi$, $\psi$ since the association rules were introduced in [1]. A GUHA procedure ASSOC [5] mines for such association rules. It was implemented several times [5, 6].

The boom of association rules in the 1990s was the start of a new effort in the study of association rules $\varphi \approx \psi$. The new results can be understood as a logic of association rules [7]. The procedure 4ft-Miner – a new enhanced

implementation of the ASSOC procedure has been developed and a research on automation of data mining with association rules and domain knowledge has been initiated [8, 9]. For more information see papers cited in [5, 7, 8].

The goals of the talk are:

- to introduce basic features of association rules related to market basket analysis
- to present an introduction to the GUHA method and related association rules
- to show examples of applications of the GUHA procedure 4ft-Miner to real data
- to introduce possibilities of automation of dealing with domain knowledge in data mining with association rules
- to present related theoretical results concerning logic of association rules.

# References


[1] Agrawal R., Srikant R. (1994) Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference Santiago, Chile, 1994

[2] Geng L., Hamilton H.J. (2006) Interestingness Measures for Data Mining: A survey. ACM Comput. Surv. **38**, 1-32

[3] Hájek P., Havel I., Chytil M. (1966) The GUHA method of automatic hypotheses determination, Computing 1 293-308.

[4] Hájek, P., Havránek, T. (1978) Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Springer, Berlin Heidelberg New York

[5] Hájek, P., Holeňa, M., Rauch, J. (2010) The GUHA method and its meaning for data mining. J. Comput. Syst. Sci. **76**, 34–48

[6] Ralbovský, M., Kuchař, T. (2009) Using Disjunctions in Association Mining. In: Perner, P. (ed.) Advances in Data Mining, pp. 339-351. Springer, Heidelberg

[7] Rauch, J. (2013) Observational Calculi and Association Rules. Berlin : Springer-Verlag, 296 pp.

[8] Rauch, J. (2015) Formal Framework for Data Mining with Association Rules and Domain Knowledge – Overview of an Approach. Fundamenta Informaticae, **137** No 2, pp. 1–47

[9] Šimůnek, M. (2014) LISp-Miner Control Language – description of scripting language implementation. Journal of System Integration, **5** No 2, `http://www.si-journal.org/index.php/JSI/article/view/193`