ALGEBRAIC METHODS OF DATA RETRIEVAL

K.Adaricheva[1]*, J.B.Nation[2]

[1]School of Science and Technology, Nazarbayev University, Astana, Kazakhstan; *kira.adaricheva@nu.edu.kz; [2]University of Hawaii, Honolulu, USA

INTRODUCTION.

Discovery of (strong) *association rules*, or implications, between the attributes of a binary table is an important task in data management, and it finds application in artificial intelligence, data mining and the semantic web. The challenge of the task is to deal with the large data sets, when the most known approaches require a time that depends exponentially on the size of the data [1,2].

RESULTS AND DISCUSSION.

We introduce a novel approach for the discovery of a specific set of implications, called the D-basis [3], that provides a representation for a reduced binary table, based on the structure of its Galois lattice. At the core of the new algorithm is the D-relation defined in the lattice theory framework, and the hypergraph dualization that allows us to effectively produce the set of transversals for a given Sperner hypergraph. The latter algorithm, developed by specialists from Rutgers Center for Operations Research [4], has already found numerous applications in solving the optimization problems in data base theory, artificial intelligence and game theory. The strength of the algorithm is in its theoretical quasi-polynomial upper bound, while it scales almost linear in practical testing. One application of the method is in analysis of gene expression data related to a particular phenotypic variable. Some initial testing is done for data provided by the University of Hawaii Cancer Center.

CONCLUSIONS.

The first tests with the data gives confirmation of advantages of a new approach over existing methods of data analysis, involving the structure of Galois lattice or others, developed in data mining field.

ACKNOWLEDGMENTS.

The work on this project was supported by seed grant N 13/42 (2013-2015) awarded by Nazarbayev University. The algorithm programming code was created by undergraduate students of Yeshiva University (New York) T. Moldwin and J. Blumenkopf. The code incorporated the call to existing implementation of hypergraph dualization algorithm by T. Uno [5]. In the testing we were assisted by Prof. U. Norbisrath (CS SST) and A. Amanbekkyzy (Math Dept SST). Data on 24 cancer patients was provided by Director of Laboratory of Bioinformatics, University of Hawaii Cancer Center, Dr. Gordon S. Okimoto.

REFERENCES.

1. R. Agrawal, R.Sricant, H.Mannila, H.Toivonen, A.Verkamo (1996). Fast discover of association rules, in Advances in Knowledge Discovery, U.Fayyad et al. eds., MIT Press.
2. M. Kryszkiewicz. (2002). Concise representation of frequent patterns and association rules, Habil. Thesis, Institute of Computer Science, Warsaw University of Technology, Poland.
3. K. Adaricheva, J.B. Nation, R. Rand. (2013). Ordered direct implicational basis of a finite closure system, Disc. Appl. Math., 161: 707—723.
4. E. Boros, K. Elbassioni, V. Gurvich, L. Khachiyan. (2002). Generating dual-bounded hypergraphs, Optimization Methods and Software, 17: 749—781.
5. K. Murakami, T.Uno. (2013). Efficient algorithms for dualizing large scale hypergraphs, Proceedings of ALENEX'13, P. Sanders and N.Zeh, eds., New Orleans, Louisiana, pp.1-13.