

AT THE CROSSROADS OF FREQUENT CLOSED ITEMSET BASED ALGORITHMS
AND GENERIC BASES OF ASSOCIATION RULES: ACTUAL PERFORMANCES AND
CHALLENGES

Proposed by: S. BEN YAHIA
Faculty of Sciences of Tunis
Department of Computer Science,
Campus Universitaire, 1060, Tunis, Tunisia
Phone: + 216 98 214 650, Fax: + 216 71 885 190
e-mail: `sadok.benyahia@fst.rnu.tn`

Description

The last two decades witnessed an explosive progress in networking, storage, and processing technologies resulting in an unprecedented amount of digitization of data. As a side effect, classical retrieval tools proved to be unable to go further beyond the top of the Iceberg. Indeed, there was an important need for tools or techniques to delve and efficiently discover valuable, non-obvious information from large databases. Data Mining, with a clear promise to do so, is the discovery of hidden information found in databases and can be viewed as a step in the Knowledge Discovery in Databases (KDD) process. Much research in Data Mining has focused on the discovery of association rules from large databases. As a side effect, exploiting and visualizing association rules became far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a dataset. This fact bootstrapped the development of more acute techniques or methods to reduce the size of the reported rule sets. In this context, the battery of results provided by the Formal Concept Analysis (FCA) permitted to define "irreducible" nucleus of association rule subsets better known as generic bases. These bases constitute reduced sets of informative rules allowing preserving the most relevant rules, without loss of information. The generation of these informative association rules relies on the extraction of frequent closed itemsets (FCI), their associated minimal generators and the underlying partial order.

This introductory tutorial mainly targets, even though not restricted to, an audience composed of advanced undergraduate and graduate students, as well as attendees from industry. Thus, most prominent approaches or algorithms are carefully sketched through toy examples. The proposed tutorial can roughly be split into two main parts:

- **Extraction of generic bases of association rules:** we shed light on generic bases of association rules. We discuss the construction approaches of these generic bases (*e.g.*, those of Guigues-Duquenne, Zaki, Kryszkiewicz, Gasmi *et al.*). We draw attention on the semantics of these generic bases towards finding an answer to the following question: "Do these generic bases provide an added-value knowledge to the end-user"?
- **FCI based algorithms: principles, data structures and actual performances:** After a detailed description of the guiding lines of the FCI based algorithms, we present a structural and analytic comparative study of these algorithms. We introduce some features (or dimensions) allowing highlighting major differences among the most prominent FCI based algorithms for mining association rules (current and future). Actual performances of these algorithms are assessed and compared. The proposed analytic comparison, in this tutorial, goes beyond those proposed by Zheng *et al.* [17], in which only sparse datasets were of interest, and Goethals and Zaki [4], where only performance curves are showed. Indeed, we try not only to show performance curves, but also to explain these performances

based on advantages and/or drawbacks of optimization strategies used in these algorithms. To obtain an in depth insight, we also present an assessment of the memory consumption of the surveyed algorithms in conjunction with the evolution of gathered information in main memory during the mining process.

Specific goals and objectives

Literature witnessed the proposal of more than one hundred of algorithms dedicated to "efficiently" extract association rules. The survival of the association rule extraction technique is up to showing its usefulness and avoiding end-user knowledge overwhelming. However, the impressive and not exploitable number of association rules—drawn from even reasonably sized contexts—is far from encouraging users to further rely on this kind of knowledge. More than a decade after the publication of the Apriori algorithm [1], the frenzy race towards more acute algorithmic performances will lead to user disappointment while neglecting the main objective: to extract a reliable knowledge, of exploitable size for the end-users.

Motivation behind such proposed stuff is that such a tutorial comes to fill a state-of-the-art gap, since up to our knowledge presenting FCI based algorithms from the lossless knowledge reduction point of view was not previously presented in well known Data Mining conferences, such as PAKDD. Hence, after more than a decade of its appearance, a meditation break for this teenager field is compellingly a must. At least, it hampers abiding towards mirage of jewels extraction that association rule technique hopes to locate.

Detailed outline

- **Formal Concept Analysis (FCA) and Knowledge Discovery in Databases (KDD): A natural connection.**
- **Extraction of generic bases of association rules:** The extraction of generic bases of association rules is of primary importance. Thus, we propose to thoroughly survey the proposals given hereafter. In addition, for each proposal, we present the guiding lines of the generic association rule extraction algorithm. For each proposal, we check the soundness of the following properties: (i) Informativity: ability to faithfully generate all redundant association rules; (ii) Derivability: validity and completeness of the inference system.
 - State-of-the-art proposals:
 - * Guigues-Duquenne basis (Guigues-Duquenne, 1986) [5],
 - * Proper basis (Luxenburger, 1991) [9],
 - * Representative Rules (Kryszkiewicz, 1998) [7],
 - * Non-Redundant Rules (Zaki, 2000) [15],
 - * Generic basis of exact rules and Informative basis of approximate rules (Bastide *et al.*, 2000) [2],
 - * Informative generic basis (Gasmi *et al.*, 2005) [3].
 - Generic association rule: study of the evolution of rule set cardinalities.
 - Towards an end-user added value: revisiting generic association rule semantics.
- **FCI based algorithms:** FCI based algorithms can be roughly split into four categories, namely "Test-and-generate", "Divide-and-conquer", "Hybrid" and "Hybrid without duplication" techniques. Based on this classification, an analytical comparison of the FCI based algorithms is presented. Aiming to stand beyond classical performance analysis, we intend

in this tutorial to provide a focal point on performance analysis based on both memory consumption and advantages and/or drawbacks of optimization strategies used in the FCI based algorithms.

- State-of-the-art algorithms:
 - * "Test-and-generate": CLOSE (Pasquier *et al.*, 1999) [10], A-CLOSE (Pasquier *et al.*, 1999) [11], TITANIC (Stumme *et al.*, 2002) [13], PRINCE (Hamrouni *et al.*, 2005) [6],
 - * "Divide-and-conquer": CLOSET (Pei *et al.*, 2000) [12],
 - * "Hybrid": CHARM (Zaki, 2002) [16],
 - * "Hybrid without duplication": LCM (Uno *et al.*, 2004) [14], DCI-CLOSED (Lucchese *et al.*, 2004) [8].
- Utilized data structures.
- Analytic study on benchmarking datasets (sparse/dense) and worst case datasets:
 - * Algorithm performances,
 - * Algorithm memory consumption.
- **Conclusions and Challenges:** The main report that can be drawn from such tutorial sheds light on the "obsessional" algorithmic effort to reduce the computation time of the interesting itemset extraction step. Thus, almost all these algorithms were focused on enumerating closed itemsets, presenting a frequency of appearance considered to be satisfactory. The dark side of this success is that such enumeration will not allow the extraction of the generic bases of association rules. After this "nightmarish" conclusion, we point out some challenges and issues that currently interest the community:
 - Towards a finer characterization of test datasets.
 - Challenging issues of mining richly structured datasets (*e.g.*, genomic datasets): soft computing techniques may be of help!
 - Well adapted visualization models issue
 - Issues towards extracting more succinct knowledge: Succinct Minimal generator, Emergent patterns, Non derivable patterns.

Expected background

Basic notions on the mathematical background of Formal Concept Analysis may be of help (but not mandatory, since such stuff will be briefly recalled during the tutorial).

Proposed length

A half-day tutorial.

Audio Visual equipment needed for the presentation

Only a data projector is of need.

Biographical sketch of the presenter

Sadok Ben Yahia obtained his Ph.D in Computer Sciences from the Faculty of Sciences of Tunis in September 2001. Since October 2002, he is an Assistant-Professor at the Computer Sciences department at the Faculty of Sciences of Tunis. He is leading a small group of researchers in Tunis, whose research interests include efficient extraction of informative and compact covers of association rules, visualization of association rules and soft computing. He is an external reviewer of well known data management journals, *i.e.*, VLDB, DMKD. Currently, he is co-guest editor of ARIMA/SACJ journals joint special issue on "Advances on end-user Data Mining techniques". By October 2006, he is organizing in Tunis and co-chairing the Program Committee of the 4th edition of the conference on Concept Lattice and its Applications (CLA'06).

References

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th Intl. Conference on Very Large Databases, Santiago, Chile*, pages 478–499, June 1994.
- [2] Y. Bastide, N. Pasquier, R. Taouil, L. Lakhal, and G. Stumme. Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the International Conference DOOD'2000, LNAI, volume 1861, Springer-Verlag, London, UK*, pages 972–986, July 2000.
- [3] Gh. Gasmi, S. BenYahia, E. Mephu Nguifo, and Y. Slimani. *TGB*: A new informative generic base of association rules. In *Proceedings of the Intl. Ninth Pacific-Asia Conference on Knowledge Data Discovery (PAKDD'05), LNAI 3518, Hanoi, Vietnam*, pages 81–90. Springer-Verlag, May 2005.
- [4] B. Goethals and M. J. Zaki. FIMI'03: Workshop on frequent itemset mining implementations. In B. Goethals and M. J. Zaki, editors, *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, volume 90 of *CEUR Workshop Proceedings, Melbourne, Florida, USA*, 19 November 2003.
- [5] J. L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 24(95):5–18, 1986.
- [6] T. Hamrouni, S. BenYahia, and Y. Slimani. PRINCE: An algorithm for generating rule bases without closure computations. In A. Min Tjoa and J. Trujillo, editors, *Proceedings of 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Springer-Verlag, LNCS 3589, Copenhagen, Denmark*, pages 346–355, 22-26 August 2005.
- [7] M. Kryszkiewicz. Representative association rules and minimum condition maximum consequence association rules. In *Proceedings of Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD), 1998, LNCS, volume 1510, Springer-Verlag, Nantes, France*, pages 361–369, 1998.
- [8] C. Lucchesse, S. Orlando, and R. Perego. DCI-CLOSED: a fast and memory efficient algorithm to mine frequent closed itemsets. In B. Goethals, M. J. Zaki, and R. Bayardo, editors, *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'04)*, volume 126 of *CEUR Workshop Proceedings, Brighton, UK*, 1 November 2004.
- [9] M. Luxenburger. Implication partielles dans un contexte. *Mathématiques et Sciences Humaines*, 29(113):35–55, 1991.
- [10] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems*, 24(1):25–46, 1999.
- [11] N. Pasquier, Y. Bastide, R. Touil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In C. Beeri and P. Buneman, editors, *Proceedings of 7th International Conference on Database Theory (ICDT'99), LNCS, volume 1540, Springer-Verlag, Jerusalem, Israel*, pages 398–416, January 1999.
- [12] J. Pei, J. Han, R. Mao, S. Nishio, S. Tang, and D. Yang. CLOSET: An efficient algorithm for mining frequent closed itemsets. In *Proceedings of the ACM-SIGMOD DMKD'00, Dallas, Texas, USA*, pages 21–30, 2000.

- [13] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Journal on Knowledge and Data Engineering (KDE)*, 2(42):189–222, 2002.
- [14] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. *Journal of Discovery Science, LNAI, volume 3245*, pages 16–31, 2004.
- [15] M. J. Zaki. Generating non-redundant association rules. In *Proceedings of the 6th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA*, pages 34–43, August 2000.
- [16] M. J. Zaki and C. J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington, Virginia, USA*, pages 34–43, April 2002.
- [17] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In F. Provost and R. Srikant, editors, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press*, pages 401–406, August 2001.