Frequent Pattern Mining: A Survey

Zheng Huang (Colin)

October 27, 2003

The original motivation for searching association rules came from the need to analyze supermarket transaction data. The first algorithm to generate all frequent item sets and confident association rules was the AIS algorithm by Agrawal et al. [AIS93], which was given together with the introduction of this mining problem. Shortly after that, the algorithm was improved and renamed Apriori, by exploiting the monotonicity property of the support of itemsets and the confidence of association rules [AS94][SA96b].The same technique was independently proposed by Mannila et al. [MTV94]. In addition, Sarawag et al. [STA98] has given an implementation of 'apriori' on relational database. However, the 'apriori' approach requires multiple scans of the database to generate all the candidates, which becomes a bottleneck. Many methods have been proposed to reduce the number of scans and candidates by hashing [PCY95], sampling [Toi96], partitioning [SON95], dynamic itemset counting [BMUT97], parallel computing [AS96], etc. Thinking of 'apriori' algorithm as a 'width-first' approach, the counter part, 'depth-first' algorithm was explored later. The first 'depth-first' algorithm was the Eclat algorithm by Zaki [ZPOL97] [SHS$^+$00], after which Han et al proposed the well known FP-growth algorithm [HPY00], which 'grows' long patterns from short ones using local frequent items without generating candidates. Agrawal et al. also proposed their depth first algorithm [AAP00] and a lexicographic tree structure [AAP01] for frequent itemset mining and long pattern generation.

Besides the above basic frequent pattern mining algorithms, many extensions have also been extensively studied. [SA95][HF95] extended the flat single dimension rule mining to multi-dimension multi-level rule mining. [SA96a][KLKF98] studied the case when attribute is quantitative instead of categorical. Max pattern [Bay98] and closed pattern [PBTL99][PHM00] are considered to be of most interest in some cases [Zak99]. Apart from basic association rules, we are often interested in causal structures [SBMU98], interesting rules [KMR+94], correlations [BMS97], emerging patterns (a pattern frequent in one class of data but infrequent in others) [DL99], contrast pattern [BP01], etc. Considering the forbidding number of possible patterns in a large database, finding all the patterns and then filtering is usually unrealistic. Noticing that many constraints are monotone or convertibly monotone, [GLW00, FH95, LNHP99, NLHP98, PHL01, JP00] proposed several algorithms to push the constraints as deeply as possible down to the frequent pattern computation. Parallel algorithms are proposed as another way to tackle the high computational cost [AS96, HKK97]. There is also some significant work on integrating frequent pattern mining into other data mining methods. [LSW97] studied association rule clustering. [LHM98] integrated classification and association rule mining.

Another popular field closely related to basic association rule mining is sequential pattern mining. The basic algorithms in this field inherited a lot from those for association rule mining and also fall into two big categories: 'Apriori-like' and 'FP-growth-like'(a.k.a. projection based). The first three 'Apriori-like' sequential pattern mining algorithms (AprioriAll, AprioriSome, DynamicSome) were proposed by Agrawal [AS95] followed by an influential apriori-based algorithm, GSP [SA96b] which takes into account the time constraints and taxonomy knowledge. Just like in the association rule mining, the apriori based approach is still limited by the high cost of multiple database scans and candidate generation. Thereby, the projection based algorithms, FreeSpan and PrefixSpan, were developed afterwards and became popular. FreeSpan [HPMA+00] recursively projected a sequence database into a set of smaller databases based on the current set of frequent patterns and mine each projected database to find its patterns. PrefixSpan [PHMA+] improved FreeSpan by using prefix-based projection. Other variations like SPADE [Zak01] (a lattice based algorithm), MEMISP [LL02] (a memory indexing based approach), SPIRIT [GRS99] (integrates constraints by using regular expression) were also proposed. Two important extensions to the basic frequent sequence mining are intensively studied: incremental mining [ZKDWC02][PZOD99] and periodic pattern analysis [HDY99].

References

[AAP00] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. *ACM SIGKDD*, 2000.

[AAP01] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *Journal of Parallel and Distributed Computing*, 61(3):350–371, 2001.

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[AS95] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[AS96] R. Agrawal and J. C. Shafer. Parallel mining of association rules. *Ieee Trans. On Knowledge And Data Engineering*, 8:962–969, 1996.

[Bay98] Roberto J. Bayardo. Efficiently mining long patterns from databases. In Laura M. Haas and Ashutosh Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 85–93. ACM Press, 1998.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. pages 265–276, 1997.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 255–264. ACM Press, 05 1997.

[BP01] S.D. Bay and M. J. Pazzani. Detecting group differnces : Mining contrast sets. In *KDD*, 2001.

[DL99] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Knowledge Discovery and Data Mining*, pages 43–52, 1999.

[FH95] Yongjian Fu and Jiawei Han. Meta-rule-guided mining of association rules in relational databases. In *KDOOD/TDOOD*, pages 39–46, 1995.

[GLW00] Gosta Grahne, Laks V. S. Lakshmanan, and Xiaohong Wang. Efficient mining of constrained correlated sets. In *ICDE*, pages 512–521, 2000.

[GRS99] Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Spirit: Sequential pattern mining with regular expression constraints. In *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pages 223–234. Morgan Kaufmann, 1999.

[HDY99] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *Fifteenth International Conference on Data Engineering*, pages 106–115, Sydney, Australia, 1999. IEEE Computer Society.

[HF95] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95), Zürich, Switzerland, September 1995*, pages 420–431, 1995.

[HKK97] Eui-Hong Han, George Karypis, and Vipin Kumar. Scalable parallel data mining for association rules. pages 277–288, 1997.

[HPMA$^+$00] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. Freespan: frequent pattern-projected sequential pattern mining. *Proc. ACM SIGKDD*, 2000.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey Naughton, and Philip A. Bernstein, editors, *2000 ACM SIGMOD Intl. Conference on Management of Data*, pages 1–12. ACM Press, 05 2000.

[JP00] Jiawei Han Jian Pei. Can we push more constrints into frequent pattern mining. 2000.

[KLKF98] Flip Korn, Alexandros Labrinidis, Yannis Kotidis, and Christos Faloutsos. Ratio rules: A new paradigm for fast, quantifiable data mining. In Ashish Gupta, Oded Shmueli, and Jennifer Widom, editors, *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 582–593. Morgan Kaufmann, 1998.

[KMR+94] Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. In Nabil R. Adam, Bharat K. Bhargava, and Yelena Yesha, editors, *Third International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407. ACM Press, 1994.

[LHM98] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pages 80–86, 1998.

[LL02] Ming-Yen Lin and Suh-Yin Lee. Fast discovery of sequential patterns by memory indexing. In *Proc. of 2002 DaWaK*, 2002.

[LNHP99] Laks V. S. Lakshmanan, Raymond Ng, Jiawei Han, and Alex Pang. Optimization of constrained frequent set queries with 2-variable constraints. pages 157–168, 1999.

[LSW97] Brian Lent, Arun N. Swami, and Jennifer Widom. Clustering association rules. In *ICDE*, pages 220–231, 1997.

[MTV94] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Efficient algorithms for discovering association rules. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

[NLHP98] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang. Exploratory mining and pruning optimizations of constrained associations rules. pages 13–24, 1998.

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. *Lecture Notes in Computer Science*, 1540:398–416, 1999.

[PCY95] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. An effective hash based algorithm for mining association rules. In Michael J. Carey and Donovan A. Schneider, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 175–186, San Jose, California, 22–25 1995.

[PHL01] J. Pei, J. Han, and L.V.S. Lakshmannan. Mining frequent itemsets with convertible constraints. In *ICDE*, 2001.

[PHM00] Jian Pei, Jiawei Han, and Runying Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.

[PHMA+] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. pages 215–226.

[PZOD99] Srinivasan Parthasarathy, Mohammed Javeed Zaki, Mitsunori Ogihara, and Sandhya Dwarkadas. Incremental and interactive sequence mining. In *CIKM*, pages 251–258, 1999.

[SA95] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. *VLDB*, 13(2–3):161–180, 1995.

[SA96a] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 1–12, Montreal, Quebec, Canada, 4–6 1996.

[SA96b] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In Peter M. G. Apers, Mokrane Bouzeghoub, and Georges Gardarin, editors, *Proc. 5th Int. Conf. Extending Database Technology, EDBT*, volume 1057, pages 3–17. Springer-Verlag, 25–29 1996.

[SBMU98] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeffrey D. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2/3):163–192, 1998.

[SHS+00] Pradeep Shenoy, Jayant R. Haritsa, S. Sudarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging vertical mining of large databases. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, volume 29, pages 22–33. ACM, 2000.

[SON95] Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In *The VLDB Journal*, pages 432–444, 1995.

[STA98] Sunita Sarawagi, Shiby Thomas, and Rakesh Agrawal. Integrating mining with relational database systems: Alternatives and implications. In Laura M. Haas and Ashutosh Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 343–354. ACM Press, 1998.

[Toi96] Hannu Toivonen. Sampling large databases for association rules. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, *In Proc. 1996 Int. Conf. Very Large Data Bases*, pages 134–145. Morgan Kaufman, 09 1996.

[Zak99] Mohammed Javeed Zaki. Charm: An efficient algorithm for closed association rule mining. *TR99-10, Rensselaer Polytechnic Institute*, 1999.

[Zak01] Mohammed Javeed Zaki. Spade: An efficient algorithm for mining frequent sequences. In *Machine Learning*, 2001.

[ZKDWC02] M. Zhang, B. Kao, and Chi Lap Yip David WaiLok Cheung. Efficient algorithms for incremental update of frequent sequences. In *PAKDD*, 2002.

[ZPOL97] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New algorithms for fast discovery of association rules. In David Heckerman, Heikki Mannila, Daryl Pregibon, Ramasamy Uthurusamy, and Menlo Park, editors, *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, pages 283–296. AAAI Press, 12–15 1997.