Chapter 7

Advanced Pattern Mining

7.1 Bibliographic Notes

This chapter described various ways in which the basic techniques of frequent itemset mining (presented in Chapter 6) have been extended. One line of extension is mining multilevel and multidimensional association rules. Multilevel association mining was studied in Srikant and Agrawal [SA95], and Han and Fu [HF95]. In Srikant and Agrawal [SA95], such mining was studied in the context of generalized association rules, and an R-interest measure was proposed for removing redundant rules. Mining multidimensional association rules using static discretization of quantitative attributes and data cubes was studied by Kamber, Han, and Chiang [KHC97]. Another line of extension is to mine patterns on numeric attributes. Srikant and Agrawal [SA96] proposed a non-grid-based technique for mining quantitative association rules, which uses a measure of partial completeness. Mining quantitative association rules based on rule clustering was proposed by Lent, Swami, and Widom [LSW97]. Techniques for mining quantitative rules based on x-monotone and rectilinear regions were presented by Fukuda, Morimoto, Morishita, and Tokuyama [FMMT96], and Yoda, Fukuda, Morimoto, et al. [YFM+97]. Mining (distance-based) association rules over interval data was proposed by Miller and Yang [MY97]. Aumann and Lindell [AL99] studied the mining of quantitative association rules based on a statistical theory to present only those rules that deviate substantially from normal data.

Mining rare patterns by pushing group-based constraints was proposed by Wang, He and Han [WHH00]. Mining negative association rules was discussed by Savasere, Omiecinski and Navathe [SON98], and by Tan, Steinbach and Kumar [TSK05].

Constraint-based mining directs the mining process towards patterns that are likely of interest to the user. The use of metarules as syntactic or semantic filters defining the form of interesting single-dimensional association rules was proposed in Klemettinen, Mannila, Ronkainen, et al. [KMR+94]. Metarule-
CHAPTER 7. ADVANCED PATTERN MINING

guided mining, where the metarule consequent specifies an action (such as Bayesian clustering or plotting) to be applied to the data satisfying the metarule antecedent, was proposed in Shen, Ong, Mitbander, and Zaniolo [SOMZ96]. A relation-based approach to metarule-guided mining of association rules was studied in Fu and Han [FH95]. Methods for constraint-based mining using pattern pruning constraints were studied by Ng, Lakshmanan, Han, and Pang [NLHP98], Lakshmanan, Ng, Han, and Pang [LNHP99], and Pei, Han, and Lakshmanan [PHL01]. Constraint-based pattern mining by data reduction using data pruning constraints was studied by Bouchi, Giannotti, Mazzanti, and Pedreschi [BGMP03], and Zhu, Yan, Han and Yu [ZYHY07]. An efficient method for mining constrained correlated sets was given in Grahne, Lakshmanan, and Wang [GLW00]. A dual mining approach was proposed by Bucila, Gehrke, Kifer, and White [BGKW03]. Other ideas involving the use of templates or predicate constraints in mining have been discussed in [AK93], [DT93], [HK91], [LHC97], [ST96], [SVA97].

Traditional pattern mining methods encounter challenges when mining high-dimensional patterns, with applications like bioinformatics. Pan, Cong, Tung, et al. [PCT+03] proposed CARPENTER, a method for finding closed patterns in high-dimensional biological datasets, which integrates the advantages of vertical data formats and pattern-growth method. Pan, Tung, Cong, and Xu [PTCX04] proposed COBBLER, which finds frequent closed itemsets by integrating row enumeration with column enumeration. Liu, Han, Xin, and Shao [LHXS06] proposed TDClose to mine frequent closed patterns in high-dimensional data by starting from the maximal rowset, integrated with a row-enumeration tree. It uses the pruning power of the minimum support threshold to reduce the search space. For mining rather long patterns, called colossal patterns, Zhu, Yan, Han, et al. [ZYH+07] developed a core pattern fusion method that leaps over an exponential number of intermediate patterns to reach colossal patterns.

To generate a reduced set of patterns, recent studies have focussed on mining compressed sets of frequent patterns. Closed patterns can be viewed as a lossless compression of frequent patterns, whereas maximal patterns can be viewed as a simple lossy compression of frequent patterns. Top-k patterns, such as that by Wang, Han, Lu, and Tsvetkov [WHLT05], and error-tolerant patterns by Yang, Fayyad, and Bradley [YFB01] are alternative forms of interesting patterns. Afrati, Gionis, and Mannila [AGM04] proposed to use $k$ itemsets to cover a collection of frequent itemsets. For frequent itemset compression, Yan, Cheng, Han, and Xin [YCHX05] proposed a profile-based approach, and Xin, Han, Yan, and Cheng [XHYC05] proposed a clustering-based approach. By taking into consideration of both pattern significance and pattern redundancy, Xin, Cheng, Yan, and Han [XCYH06] proposed a method for extracting redundancy-aware top-k patterns.

Automated semantic annotation of frequent patterns is useful for explaining the meaning of patterns. Mei, Xin, Cheng, et al. [MXC+07] studied methods for semantic annotation of frequent patterns.

An important extension to frequent itemset mining is mining sequence and structural data. This includes mining sequential patterns (such as in Agrawal
and Srikant [AS95], Pei, Han, Mortazavi-Asl et al. [PHMA+01, PHMA+04],
and Zaki [Zak01]), mining frequent episodes (Mannila, Toivonen, and Verkamo
[MTV97]), mining structural patterns (e.g., Inokuchi, Washio, and Motoda
[IWM98], Karamochi and Karypis [KK01], and Yan and Han [YH02]), mining
cyclic association rules (Özden, Ramaswamy, and Silberschatz [ORS98]), inter-
transaction association rule mining (Lu, Han, and Feng [LHF98]), and calen-
dric market basket analysis (Ramaswamy, Mahajan, and Silberschatz [RMS98]).
Mining such patterns will be studied in-depth in the second volume of this book.

Pattern mining has been extended to help effective data classification and
clustering. Pattern-based classification (such as Liu, Hsu and Ma [LHM98],
and Cheng, Yan, Han, and Hsu [CYHH07] is discussed in Chapter 9. Pattern-
based cluster analysis (such as Agrawal, Gehrke, Gunopulos, and Raghavan
[AGGR98], and H. Wang, W. Wang, Yang, and Yu [WWYY02]) is discussed in
Chapter 11.

Pattern mining also helps many other data analysis and processing tasks,
such as cube gradient mining and discriminative analysis (Imielinski, Khachiyan
and Abdorghani [IKA02]; Dong, Han, Lam, et al. [DHL+04]; Ji, Bailey and Dong
[JBD05]), discriminative pattern-based indexing (Yan, Yu and Han [YYH05]),
and discriminative pattern-based similarity search (Yan, Zhu, Yu and Han
[YZYH06]).

Pattern mining has been extended to mining spatial, temporal, time-series,
multimedia data, and data streams. Mining spatial association rules or spatial
collocation rules was studied by Kopierski and Han [KH95], Xiong, Shekhar,
Huang, et al. [XSH+04], and Cao, Mamoulis and Cheung [CMC05]. Pattern-
based mining of time-series data is discussed in Shieh and Keogh [SK08] and
Ye and Keogh [YK09]. There are many studies on pattern-based mining of
multimedia data, such as Zaïane, Han and Zhu [ZHZ00], and Yuan, Wu and
Yang [YWY07]. Methods for mining frequent patterns on stream data have
been proposed by many researchers, including Manku and Motwani [MM02],
Karp, Papadimitriou and Shenker [KPS03], and Metwally, Agrawal, and El
Abbadi [MAA05]. These pattern mining methods will be discussed in-depth in
the second volume of this book.

Pattern mining has broad applications. Application areas include computer
science, such as software bug analysis, sensor network mining, and performance
improvement of operating systems. For example, CPMiner by Li, Lu, Myag-
mar, and Zhou [LLM204] uses pattern mining to identify copy-pasted code for
bug isolation. PRMiner by Li and Zhou [LZ05] uses pattern mining to extract
application-specific programming rules from source code. Discriminative pat-
tern mining is used for program failure detection to classify software behaviors
(Lo, Cheng, Han, et al. [LCH+09]), and for troubleshooting in sensor networks
(Khan, Le, Ahmadi et al. [KLA+08]). Such applications will also be covered in
the second volume of this book.
CHAPTER 7. ADVANCED PATTERN MINING
Bibliography


[LHXS06] H. Liu, J. Han, D. Xin, and Z. Shao. Mining frequent patterns on very high dimensional data: A top-down row enumeration approach. In Proc. 2006 SIAM Int. Conf. Data Mining (SDM’06), Bethesda, MD, April 2006.


