Pattern Discovery in Temporal Databases: Some Recent Results

Alexander Tuzhilin

Information Systems Department
Stern School of Business
New York University
atuzhili@stern.nyu.edu

My recent work on pattern discovery in temporal databases falls into two broad categories: (1) the use of temporal logic for the specification and discovery of temporal patterns and (2) discovery of temporal rules in Web logfiles. These two topics are presented below.

1. The use of temporal logic for the specification and discovery of temporal patterns

In their influential paper [MTV95], Mannila et al. formulated the problem of discovering frequently occurring temporal patterns in sequences, where temporal patterns are specified using the notion of an *episode* [MTV95]. It has been recognized by Padmanabhan and Tuzhilin [PT96] that temporal patterns can be specified using temporal logic and that this method generalizes the episodes approach proposed in [MTV95] in two ways. First, temporal logic is more expressive than the episodes formalism. Therefore, it allows to specify a richer set of temporal patterns. Second, the approach proposed in [PT96] is applicable not only to sequences (i.e., propositional case), but also to temporal databases (the first-order case). It is also described in [PT96] how frequent occurrences of patterns expressed with temporal logic can be discovered using temporal logic programming methods [AM89]. In particular, [PT96] uses an observation that any temporal logic formula can be simulated with a temporal logic program. Therefore, in order to find frequently occurring temporal patterns belonging to a certain class of temporal logic formulas, one has to generate an appropriate temporal logic program that simulates these formulas and counts the number of occurrences of these patterns.

Although frequency of patterns is an important and popular measure of their "interestingness," it turns out that in certain applications, such as intrusion detection and Web-based applications, an alternative measure based on "unexpectedness" is more important than frequency [BT98a]. Berger and Tuzhilin [BT98a] introduce a probabilistic measure of interestingness based on *unexpectedness,* whereby a pattern *P* is interesting if the ratio of the *actual* number of occurrences of *P* significantly deviates from the *expected* number of occurrences of *P*. One of the main problems with this measure of interestingness is that it is not monotonic, i.e. an addition of an event to an uninteresting pattern *can* result in the discovery of an interesting pattern, unlike the case with the frequency measure [MTV95]. This property makes the problem of discovering unexpected patterns hard (it is shown in [BT98a] that it is an NP-hard problem). [BT98a] addresses this problem by developing an efficient algorithm that discovers most of the unexpected patterns (but not all of them) for a fragment of a propositional linear temporal logic. This algorithm keeps track of the portions of temporal patterns and uses a greedy heuristic to determine the most promising pattern to expand. This process continues until all the partial patterns are expanded and all the resulting unexpected patterns are discovered. Because of the non-monotonicity property, the algorithm does not guarantee the discovery of all the unexpected patterns. However, as the experiments show, it discovers a large percentage of all the interesting patterns (between 92.4% and 98.4% in our experiments). This algorithm was further improved in [BT98b], mainly, by reducing the space required to store some intermediate results.

The algorithms presented in [BT98a, BT98b] have been experimentally tested on several applications, including Web-based and intrusion detection applications. In the Web application, the proposed algorithms searched for various unexpected traversal patterns at a Web site. In our experiments we used two Web sites: one being artificially created and another being a major Web site at a large university. The Web application will be described further below. In case of an intrusion detection application, unexpected patterns indicate regions of suspicious activities which could be indicators of possible intrusions. The algorithms presented in [BT98a, BT98b] were used on the data containing traces of *sendmail* system calls collected at the University of New Mexico [FHS+96], and several regions of potential intrusions have been detected. Unfortunately, the *sendmail* system calls data is not labeled. Therefore, we could not test our results for accuracy, and we plan to test them on a better set of data that should be provided shortly to the intrusion detection community.

2. Pattern discovery in Web logfiles

Web logfiles constitute a very interesting example of a temporal database, and we worked on the problem of pattern discovery in Web logfiles. In particular, we considered two types of patterns for this Web application: patterns as sequences (as explained in Section 1, e.g. "before a user comes to page X, he or she visits page Y immediately followed by page Z") and patterns as IF-THEN rules (e.g., IF a user visits page X before page Y THEN the user visits page Z). We describe the discovery of these two types of Web patterns now.

As explained in Section 1, we studied in [BT98a] the problem of discovering unexpected temporal patterns expressed in temporal logic. As was explained before, a pattern is *unexpected* if the actual number of its occurrences significantly deviates from the expected number of these occurrences. The actual number of occurrences can be computed directly from the Web logfile. The expected number can be estimated by knowing the link structure of the Web site and estimating conditional probabilities of moving from one page to another. As explained before, we applied the discovery algorithm presented in [BT98a] to two Web sites. The first site is an artificially created Web site. The algorithm managed to find 94.4% of all the unexpected patterns at that site. The second site is a major Web site at a large university. It contained 4459 pages with 37954 links between them. The algorithm was tested on two sequences of page accesses for two individuals, each sequence having more than 1400 events and extends over a period of 9 months. Our results were less encouraging: the algorithm found only 2 patterns of length greater than 2 for both sequences. These results are attributed to the following factors [BT98a]. First, the total number of pages was too large in comparison to the number of events. Second, the Web site structure is constantly changing, whereas [BT98a] considered the structure taken only at a specific point in time. Therefore, calculations of the expected number of patterns are only approximations of reality. Third, the string of events was treated as one very long session, whereas in reality there are many different sessions, and this causes technical problems described in [BT98a]. In summary, [BT98a] identified these as serious issues that need to be addressed in order to be able to discover more meaningful patterns in Web logfile data.

In contrast to [BT98a], Padmanabhan and Tuzhilin [PT98] addressed the problem of discovering unexpected *rules* in Weblog data. The methods described in [PT97] and [PT98] are targeted towards the discovery of association rules [AIS93] and their extensions [PT98] and use the system of user-specified (or learned) *beliefs* for discovering unexpected patterns contradicting these beliefs. In the Web application example, one may believe that "for all pages, for all weeks, the number of hits to a page each week is approximately equal to the page's average weekly hits." This belief can lead to the discovery of the following patterns using the discovery methods described in [PT97, PT98]: (1) for a certain "Call for Papers" page, in the weeks from September 10 through October 29, the weekly access count is much higher than the average; (2) for a certain faculty position advertisement page, the weeks closest to the deadline had unusually high visitation activity. It is interesting to note that the approach described in [PT97, PT98] led to the discovery of several unexpected Web visitation rules.

References

[AIS93] Agrawal, R., Imielinski, T. and Swami, A. "Mining association rules between sets of items in large databases." In *Proc. of the ACM SIGMOD Conference on Management of Data,* pp. 207-216, 1993.

[AM89] Abadi, M. and Manna, Z. "Temporal logic programming." *Journal of Symbolic Computation,* 8:277-295, 1989.

[BT98a] Berger, G. and Tuzhilin, A. "Discovering unexpected patterns in temporal data using temporal logic." In *Temporal Databases - Research and Practice.* O. Etzion, S. Jajodia, S. Sripada (eds). Springer-Verlag, 1998.

[BT98b] Berger, G. and Tuzhilin, A. "A temporal logic approach to discovering unexpected patterns in sequences." Working Paper IS-98-007, Stern School of Business, NYU, 1998.

[FHS+96] Forrest, S., Hofmeyer, A,, Somayaji, A. and Longstaff, T. "A sense of self for Unix processes." In *Proc. of the 1996 IEEE Symp. on Security and Privacy,* 1996.

[MTV95] Mannila, H., Toivonen, H. and Verkamo, A.I. "Discovering frequent episodes in sequences." In *Proc. of the 1st International Conference on Knowledge Discovery and Data Mining,* 1995.

[PT96] Padmanabhan, B. and Tuzhilin, A. "Pattern Discovery in Temporal Databases: A Temporal Logic Approach." *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining,* 1996.

[PT97] Padmanabhan, B. and Tuzhilin, A. "Discovering Unexpected Rules in Data Mining Applications." *Proc. of the Wrkshp on Information Technology and Systems,* 1997.

[PT98] Padmanabhan, B. and Tuzhilin, A. "A Belief-Driven Method for Discovering Unexpected Patterns." Submitted for publication.