Extended Abstract of a paper prepared for the Lausanne Conference on Sequence Analysis (LaCOSA), University of Lausanne, June 6-8, 2012

Exploring the sequencing and timing of life events

Gilbert Ritschard
LIVES and Institute for Demographic and Life Course Studies (IDEMO)
University of Geneva

Preliminary, please do not cite without permission, May 1, 2012

Abstract

This paper explains how data-mining-based techniques can be used for discovering interesting knowledge from sequences of life events, that is, to find out how people sequence important life events. We illustrate with data from the biographical survey conducted by the Swiss Household Panel in 2002. The focus is on the sequencing of events in the occupational life course and of events such as starting a union and childbirth that affect the living arrangement. Addressed methods include finding of frequent sequential patterns, identification of discriminant subsequences and clustering of event sequences.

Mining of sequentially organized data has been successfully exploited in many domains such as genetics, device control, speech recognition as well as for automatic text, customer behavior and web logs analysis. In the social sciences sequence exploration mainly focuses on state sequences, and consists typically in building typologies by means of the optimal matching approach (Abbott and Forrest, 1986). In contrast, we focus in this chapter on event sequences rather than state sequences. Let us clarify the difference between the two.

A state, such as being jobless, lasts the whole considered unit of time while an event, for example ending a job, occurs at a certain time point and has no duration. The event does not last, but provokes, possibly in conjunction with other events, a state change. In state sequences, the positions in the sequence reflect the duration since the beginning of the sequence, while in event sequences they just inform about the number of precedent events. Therefore, while state sequences are particularly of interest for studying durations and timing in life courses, event sequences are especially useful when the concern is the sequencing, i.e., the order in which events occur. State sequences have received a lot of attention in the social sciences, especially since the popularization of the optimal-matching-based methods by Abbott in the late 80's (Abbott and Forrest, 1986;

Aisenbrey and Fasang, 2010), and there exist nowadays efficient pieces of software to explore such state sequences (e.g., Brzinsky-Fay et al., 2006; Gabadinho et al., 2011). As stated above, we do not address methods for state sequences in this chapter, but focus on event sequences. The aim is to show what we can do with event sequences and the kind of results that we can expect from an event sequence analysis.

Although the exploration of event sequences has already been considered in the social science literature (Abbott, 1983, 1991; Blockeel et al., 2001; Billari et al., 2006), the approach followed in those works mainly consists, with the noticeable exception of Blockeel et al. (2001), in looking at the frequencies of a priori defined subsequences of interest. Here, we adopt a more holistic point of view and explore all subsequences that can be found in the data. We place our approach in the line of work developed by the data mining community (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) as an extension of the mining of frequent—non ordered—patterns.

Broadly, the addressed methods consist in finding the most frequent subsequences, i.e., the most common ways of sequencing life events, and then in finding out among them those that best discriminate between groups such as women and men for example. The measure of pairwise dissimilarities between event sequences is also addressed and we show how such dissimilarities can be used for clustering time-stamped event sequences.

For the social sciences, the exploration of event sequences should permit to answer questions such as

- What is the most typical succession of family or professional life events?
- Are there standard ways of sequencing those events?
- What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
- How is the sequencing of events related to covariates?
- Which event sequencings do best discriminate groups such as men and women?

As already mentioned, our aim is to demonstrate the scope of sequential event pattern mining in social sciences. We exploit for that the features offered by the TraMineR package (Gabadinho et al., 2011) for event sequences (for TraMineR's event sequence features, see Studer et al., 2010, Bürgin et al., 2012 as well as the User's guide, Gabadinho et al., 2009) and illustrate with data on Swiss cohabitational and occupational life courses.

References

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods 16*(4), 129–147.

Abbott, A. (1991). The order of professionalization: An empirical analysis. *Work and Occupations 18*(4), 355–384.

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Agrawal, R. and R. Srikant (1995). Mining sequential patterns. In P. S. Yu and A. L. P. Chen (Eds.), *Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan*, pp. 487–499. IEEE Computer Society.

Aisenbrey, S. and A. E. Fasang (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research 38*(3), 430–462.

Bettini, C., X. S. Wang, and S. Jajodia (1996). Testing complex temporal relationships involving multiple granularities and its application to data mining (extended abstract). In *PODS '96: Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, New York, pp. 68–78. ACM Press.

Billari, F. C., J. Fürnkranz, and A. Prskawetz (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

Blockeel, H., J. Fürnkranz, A. Prskawetz, and F. Billari (2001). Detecting temporal change in event sequences: An application to demographic data. In L. De Raedt and A. Siebes (Eds.), *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001*, Volume LNCS 2168, pp. 29–41. Freiburg in Brisgau: Springer.

Brzinsky-Fay, C., U. Kohler, and M. Luniak (2006). Sequence analysis with Stata. *The Stata Journal 6*(4), 435–460.

Bürgin, R., G. Ritschard, et E. Rousseaux (2012). Exploration graphique de données séquentielles. In *Atelier Fouille Visuelle de Données : méthologie et évaluation, EGC 2012, Bordeaux*, pp. 39–50. Association EGC.

Gabadinho, A., G. Ritschard, N. S. Müller, et M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009). Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Mannila, H., H. Toivonen, and A. I. Verkamo (1995). Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995*, pp. 210–215. AAAI Press.

Mannila, H., H. Toivonen, and A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI E-19*, 37–48.

Zaki, M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.