Chapter XXIX
Fuzzy Sequential Patterns for Quantitative Data Mining

Céline Fiot
University of Montpellier II – CNRS, France

ABSTRACT
The explosive growth of collected and stored data has generated a need for new techniques transforming these large amounts of data into useful comprehensible knowledge. Among these techniques, referred to as data mining, sequential pattern approaches handle sequence databases, extracting frequently occurring patterns related to time. Since most real-world databases consist of historical and quantitative data, some works have been done for mining the quantitative information stored within such sequence databases, uncovering fuzzy sequential patterns. In this chapter, we first introduce the various fuzzy sequential pattern approaches and the general principles they are based on. Then, we focus on a complete framework for mining fuzzy sequential patterns handling different levels of consideration of quantitative information. This framework is then applied to two real-life data sets: Web access logs and a textual database. We conclude on a discussion about future trends in fuzzy pattern mining.

INTRODUCTION
The amount of generated and collected data has been rapidly increasing in the last decades; these huge data and information collections are far outpacing our abilities to analyse, summarize, and extract knowledge. This explosive growth in stored data has generated a need for new techniques that can help in transforming these large quantities of data into useful comprehensible knowledge. These techniques, referred to as data mining, consist of automatically extracting patterns representing knowledge implicitly contained in large databases. Some of these approaches use principles of the fuzzy set theory; an introduction to such fuzzy data mining methods by Feil and Abonyi is included as a chapter of this book.
Among these data mining techniques, only few of them can handle an easy and complete processing of databases containing sequences of ordered events. The most appropriate one is sequential pattern mining, which consists of searching for frequently occurring patterns related to time or other order between records. Since many real-world databases—demographic phenomena, telecommunication records, medical monitoring, production processes—are highly time-correlated data, sequential pattern mining is useful to analyse such databases for targeted marketing, failure prediction, fraud detection, and so on. However, most studies focus on symbolic or qualitative sequential patterns, and the quantitative information also stored within a sequence database is often ignored. Several methods have been proposed to mine sequential patterns within historically stamped quantitative data, most of them based on the fuzzy set theory. Within the context of a supermarket basket analysis, such patterns would be, for instance, “60% of people purchase a lot of candies and few video games, and buy later a lot of toothpaste.” These patterns are characterized by their frequency, which is by definition the proportion of objects (customers) in the database that have recorded (bought) these sequences (of products). A survey of these fuzzy sequential pattern approaches shows that they are all based on the same principle: first, converting the quantitative data into fuzzy sets and membership degrees, then using these degrees to extract the frequent sequences.

Most of these approaches consider the fuzzy set theory as a tool to process quantitative information without the drawbacks of using crisp intervals. However, the use of fuzzy sets can bring more accurate analysis tools. More precisely, they can help in mining for gradual information while computing the frequency of sequential patterns. For this reason, these approaches have been generalized into a complete framework that formalizes and extends the fuzzy sequence mining methods. This framework is a complete, efficient, and scalable fuzzy approach for sequential pattern mining that enables the processing of quantitative data, with different levels of consideration of quantitative information. The end user is thus allowed to choose between the speed of result extraction and the accuracy of the obtained frequent patterns.

In this chapter, we present an introduction to the fundamental basis of sequential pattern mining, generally defining the concepts of sequence and frequency. Then we briefly describe the different proposals of fuzzy sequential pattern mining methods. In the following part, we introduce the framework generalizing these approaches. We detail how a sequence frequency can be assessed and how, within the context of the fuzzy set theory, this frequency can be computed in several ways. We also explain how different levels of information can be extracted using the different frequency computations.

After having theoretically detailed fuzzy sequential patterns, we show how to use this method for real-life database mining. The first application is the analysis of Web access logs. The second one consists of mining for word composition within a textual database. We finally end this chapter with a short discussion on future trends in fuzzy pattern mining.

**BACKGROUND**

Agrawal and Srikant (1995) initially introduced sequential patterns as a temporal extension of association rules, which have been adapted to fuzzy sequential patterns in order to handle quantitative data by several proposals.

**Sequential Pattern Mining**

Let \( DB \) be a set of object records where each record \( R \) consists of three information elements: an object ID, a record time stamp, and a set of binary attributes in the record. An attribute can be either present, meaning its value is 1 or true, or it can be absent, meaning its value is 0 or false. In the context of crisp sequence mining, algorithms only consider the presence of an attribute; that is, each time that