Chapter 10
Maximal Sequential Patterns: A Tool for Quantitative Semantic in Text Analysis

René Arnulfo García-Hernández
Autonomous University of the State of Mexico, Mexico

J. Fco. Martínez-Trinidad
National Institute of Astrophysics, Mexico

J. Ariel Carrasco-Ochoa
National Institute of Astrophysics, Mexico

ABSTRACT
This chapter introduces maximal sequential patterns, how to extract them, and some applications of maximal sequential patterns for document processing and web content mining. The main objective of this chapter is showing that maximal sequential patterns preserve document semantic, and therefore they could be a good alternative to the word and n-gram models. First, this chapter introduces the problem of maximal sequential pattern mining when the data are sequential chains of words. After, it defines several basic concepts and the problem of maximal sequential pattern mining in text documents. Then, it presents two algorithms proposed by the authors of this chapter for efficiently finding maximal sequential patterns in text documents. Additionally, it describes the use of maximal sequential patterns as a quantitative semantic tool for solving different problems related to document processing and web content mining. Finally, it shows some future research directions and conclusions.

INTRODUCTION
In the last years, frequent pattern mining has been an intensively studied task into datamining (Han, Cheng, Xin, & Yan, 2007). Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with a frequency not less than a user-specified threshold. Frequent pattern mining constitutes an important step for mining associations, correlations, finding interesting relationships among data, data indexing, classification, clustering, text retrieval, among other data mining tasks. Besides, frequent patterns are useful for solving complex problems about data analysis.
Maximal Sequential Patterns

Therefore, frequent pattern mining has become an important research area in datamining. Frequent pattern mining was first proposed in (Agrawal, Imieliński, & Swami, 1993) for market basket analysis, finding associations between the different items that customers place in their “shopping baskets”. Since this first proposal there have been many research publications proposing efficient mining algorithms, most of them, for mining frequent patterns in transactional databases. However, mining frequent patterns in text document databases is a problem less studied in the literature. Sequential pattern mining in text document databases has the goal of finding all the subsequences that are contained at least $\beta$ times in a collection of text documents or in a single text document, where $\beta$ is a user-specified support threshold. This discovered set of frequent sequences contains the maximal frequent sequences (from now on we will use the term Maximal Sequential Patterns, MSP), which are not a subsequence of any other frequent sequence, in this way, the MSPs are a compact representation of the whole set of frequent sequences. Therefore, in the same way as occurs in transactional databases, the sequential pattern mining in text document databases plays an important role, because it allows identifying valid, novel, potentially useful and ultimately understandable patterns. Additionally, since maximal sequential patterns can be extracted from text documents independently of the language without losing their sequential nature they can be used to solve more complex problems as question answering, authorship attribution, automatic text summarization, document clustering, text retrieval, among others. All of them related to web content mining (Kosala & Blockeel, 2000; Srivastava, 2000; Zhang & Segall, 2009).

This chapter has five goals. First, it will introduce the problem of maximal sequential pattern mining when the data are sequential chains of words, which is a little studied problem in the literature. Second, it will define several basic/generic concepts and the problem of maximal sequential pattern mining in text documents. Third, it presents two algorithms proposed by the authors of this chapter for efficiently finding MSPs in text documents. Fourth, it will describe the use of maximal sequential patterns as a quantitative semantic tool for solving different problems related to document processing and web content mining. Fifth, it will show future research directions and conclusions.

BACKGROUND

In the last decades, there has been a rapid growth of information stored in electronic devices. In (Leavitt, 2002) the author reported that about the 20 percent of the electronic information in the companies is stored in structured databases, where objects or registers are easily accessible. This situation motivated the interest for analyzing information stored in this kind of databases. A research area that focuses on the analysis of information stored in structured databases is the Knowledge Discovery defined in (Fayad, Piatetsky-Shapiro, & Padhraic, 1996) as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. These patterns should be easily understandable by the user. The key step in the process of knowledge discovery in databases is Datamining, which following to (Fayad, Piatetsky-Shapiro, & Padhraic, 1996) is “a step in the Knowledge Discovery in Databases process that consists in applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data”.

An important branch of data mining is the Sequential Pattern Mining, which finds high-frequency patterns. In (Agrawal & Srikant, 1995) the authors introduced it as follows. Given a database of transactions where each transaction is an ordered list of items, to find the maximal sequential patterns among all the patterns that have a certain user-specified minimum support, where the