A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection

René A. García-Hernández José Fco. Martínez-Trinidad
Jesús Ariel Carrasco-Ochoa

National Institute of Astrophysics, Optics and Electronics (INAOE)
Puebla, México
{renearnulfo, fmartine, ariel}@inaoep.mx

Abstract. Sequential pattern mining is an important tool for solving many data mining tasks and it has broad applications. However, only few efforts have been made to extract this kind of patterns in a textual database. Due to its broad applications in text mining problems, finding these textual patterns is important because they can be extracted from text independently of the language. Also, they are human readable patterns or descriptors of the text, which do not lose the sequential order of the words in the document. But the problem of discovering sequential patterns in a database of documents presents special characteristics which make it intractable for most of the apriori-like candidate-generation-and-test approaches. Recent studies indicate that the pattern-growth methodology could speed up the sequential pattern mining. In this paper we propose a pattern-growth based algorithm (DIMASP) to discover all the maximal sequential patterns in a document database. Furthermore, DIMASP is incremental and independent of the support threshold. Finally, we compare the performance of DIMASP against GSP, DELISP, GenPrefixSpan and cSPADE algorithms.

1. Introduction

The *Knowledge Discovery in Databases* (KDD) is defined by Fayyad [1] as "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". The key step in the knowledge discovery process is the data mining step, which following Fayyad: "consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data". This definition has been extended to *Text Mining* like: "consisting of applying text analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the text". So, text mining is the process that deals with the extraction of patterns from textual data. This definition is used by Feldman [2] to define *Knowledge Discovery in Texts* (KDT). In both KDD and KDT tasks, special attention is required in the performance of the algorithms because they are applied on a large amount of information. In particular the KDT process needs to define simple structures that can be extracted from text documents automatically and in a reasonable time. These structures must be rich enough to allow interesting KD operations [2] having in mind that in some cases the document database is updated.