The SPMF Open-Source Data Mining Library Version 2

Philippe Fournier-Viger[1](✉), Jerry Chun-Wei Lin,[2], Antonio Gomariz[3], Ted Gueniche[4], Azadeh Soltani[5], Zhihong Deng[6], and Hoang Thanh Lam[7]

[1] School of Natural Sciences and Humanities
[2] School of Computer Science and Technology
Harbin Institute of Technology Shenzhen Graduate School, China
[3] Dept. of Information and Communication Engineering, University of Murcia, Spain
[4] Dept. of Computer Science, University of Moncton, Canada
[5] Department of Computer Engineering, University of Bojnord, Bojnord, Iran
[6] School of Electronics Engineering and Computer Science, Peking University, China
[7] IBM Ireland Research Lab, Ireland
philfv8@yahoo.com, jerrylin@ieee.org, agomariz@gmail.com, ted.gueniche@gmail.com, a.soltani@ub.ac.ir, zhdeng@cis.pku.edu.cn, t.l.hoang@ie.ibm.com

Abstract. SPMF is an open-source data mining library, specialized in pattern mining, offering implementations of more than 120 data mining algorithms. It has been used in more than 310 research papers to solve applied problems in a wide range of domains from authorship attribution to restaurant recommendation. Its implementations are also commonly used as benchmarks in research papers, and it has also been integrated in several data analysis software programs. After three years of development, this paper introduces the second major revision of the library, named SPMF 2, which provides (1) more than 60 new algorithm implementations (including novel algorithms for sequence prediction), (2) an improved user interface with pattern visualization (3) a novel plug-in system, (4) improved performance, and (5) support for text mining.

Keywords: open-source library, data mining, frequent pattern mining

1 Introduction

Several open-source general purpose data mining libraries or programs have been developed such as Knime [2], Mahout [8], and Weka [9]. Although these software programs provide algorithms for many data mining tasks, they provide very few algorithms for mining frequent patterns in databases, while hundreds of algorithms have been proposed in this field during the last twenty years [**?**]. Moreover, the majority of researchers in the field of frequent pattern mining do not share their implementation or source code online. As a result, a user who wants to apply specific algorithms from this field, providing some particular features required by an application, often needs to implement the algorithms again, which

is time-consuming and requires programming knowledge. To address this issue, the SPMF (Sequential Pattern Mining Framework) [5] open-source library has been created in 2009. The goal is to provide a common library for sharing the source code of efficient implementations of frequent pattern mining algorithms to increase their use in real applications, and also to provide a set of reference implementations for researchers to compare algorithms. Initially, SPMF was designed as a library for mining frequent patterns [1] in sequences (hence its name). But over the years, it has evolved to include all kinds of pattern mining algorithms for discovering patterns such as itemsets and association rules, sequential patterns [1], periodic patterns [11], and high-utility patterns [10]. It also provides a simple user-interface for quick testing and a command-line interface for easy integration with other systems. In the past five years, SPMF has been used in more than 310 research papers to solve applied problems in a wide range of domains ranging from authorship attribution, retail forecasting, chemistry, music analysis to restaurant recommendation. The algorithm implementations of SPMF are optimized and commonly used as benchmarks in research papers. SPMF has also been integrated in several popular data analysis software programs such as ScaVis and MOA [3]. Nowadays, SPMF offers by far the largest library of pattern mining algorithms with over 120 algorithms. Moreover, it is open-source, it can be used in commercial projects, and it is an active project, unlike similar smaller projects such as Coron [4] and LUCS-PKDD [7]. Moreover, SPMF is lightweight as it has no dependencies to any other projects. The first major release of SPMF is version 0.94, released in 2013 [5]. This paper introduces the second major release of the library, named SPMF 2.

2 Novel features

SPMF 2 introduces five major novelties. First, it offers about 60 novel algorithm implementations. Thus, the number of algorithms has doubled since the previous major release, offering a greater range of algorithms to users. In particular, a novel module has been integrated in SPMF offering seven state-of-the-art algorithms for sequence prediction named DG, LZ78, AKOM, TDAG, PPM, CPT and CPT+. Sequence prediction (predicting the next symbol of a sequence of symbols based on a set of training sequences) has wide-applications in many domains such as web page prefetching and path recommendation [6]. Moreover, SPMF now offers about 20 more algorithms for utility pattern mining [10], which is probably the most active research area in frequent pattern mining. Utility pattern mining consists of finding patterns that may not be frequent but have a high-utility, where utility can be defined for example as the products generating the highest profit in a transaction database.

Second, the user interface has been improved. The main window is shown in the left side of Fig. 1. It is designed as a minimalistic user interface that let the user choose an algorithm, set its parameters and choose and input and output file, to then launch the algorithm. But an important novelty in SPMF 2 is a new pattern visualization window that let the user explore the patterns

found by any algorithm in a table view (right side of Fig. 1). Using that window, the user can browse patterns, search patterns, and apply complex filters with boolean conditions, sorts, and export the result of these operations to various formats such as text and CSV files. Thus, this window lets the user perform post-processing of the patterns found by the algorithms using various criteria.

Third, another important novel feature is a plug-in system. In SPMF 2, a user can implement new algorithms by sub-classing a class named *DescriptionOfAlgorithm*. SPMF can automatically detect algorithms sub-classing this class (which can be stored in another JAR file) and load the additional algorithms in its user-interface and show them in the same list as its built-in algorithms. This allow researchers to easily extend the software with new algorithms, and reuse the same user interface.

Fourth, in this new version of SPMF, many performance optimizations have been performed to increase the performance of the algorithms already offered in SPMF. For example, the performance of the new implementation of PrefixSpan introduced in SPMF 2 is up to 10 times faster and consumes up to twice less memory than the previous version. Extensive performance comparison of various versions of algorithms and optimizations in SPMF are not presented in this paper due to length limitations but can be found on the SPMF website at: `http://www.philippe-fournier-viger.com/spmf/`.

Fifth, support for additional input formats has been added to SPMF. In SPMF 2, mining patterns in text documents is now natively supported. Thus, algorithms for discovering patterns such as itemsets and sequential patterns can now be applied to files containing texts. This is a very important feature as SPMF has been used in many papers related to text mining but previous versions of SPMF required that the user preprocesses input files to convert them to the SPMF format, which was inconvenient. When the new version of SPMF is applied to a text document, each word is seen as a symbol, and each sentence is viewed as a transaction or sequence. The document is transformed to an internal representation used by the algorithms and the result is then transformed again to be displayed to the user.

3 Conclusion and future work

In this paper, we presented the second major release of the SPMF library (version 2), which offers many new algorithms, an improved user interface with pattern visualization, a novel plug-in system, many performance optimizations, as well as support for additional formats such as text files.

The SPMF library is an active project. Many contributors have provided algorithm implementations to the project from universities all around the world. The current development of SPMF is focused on providing more algorithms especially for discovering patterns in graphs and time-series, types of data that have not yet been considered in SPMF. Besides, an enhanced user interface for visually combining several algorithms in a workflow, and for interactive mining are currently planned for the next release.

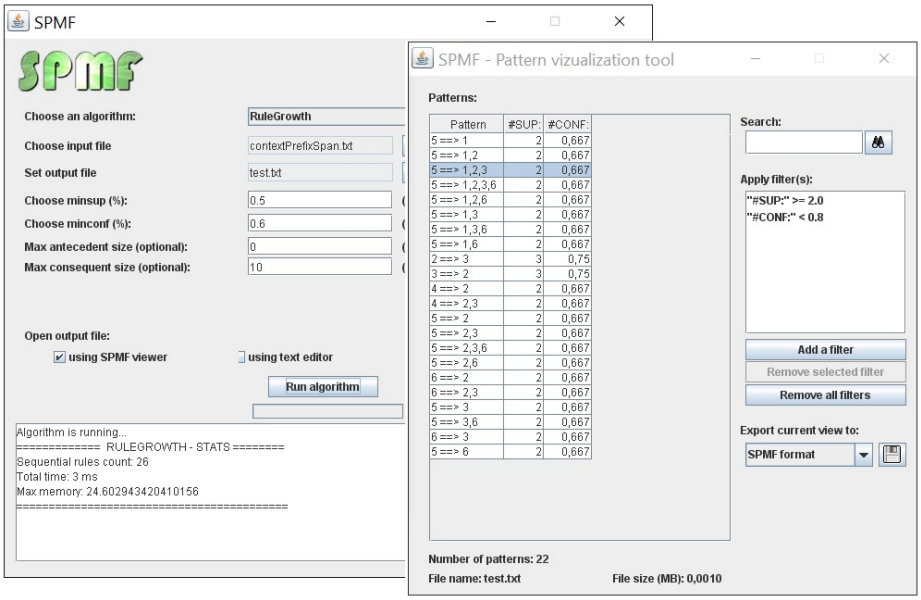

Fig. 1. The SPMF user interface

References

1. Agrawal, R., Ramakrishnan, S.: Mining sequential patterns. In: Proc. 11th Intern. Conf. Data Engineering, pp. 3–14, IEEE (1995)
2. Berthold, M. R. et al.: KNIME - the Konstanz information miner: version 2.0 and beyond. SIGKDD Explorations, 11(1): 26–31 (2009)
3. Bifet, A. et al.: MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering. Journal of Machine Learning Research (JMLR), 11, pp. 1601-1604 (2010)
4. Coron. Software available at: http://coron.loria.fr/site/index.php
5. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu., C., Tseng, V. S.: SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR), 15, pp. 3389-3393 (2014)
6. Gueniche, T., Fournier-Viger, P., Raman, R., Tseng, V. S.: CPT+: Decreasing the time/space complexity of the Compact Prediction Tree. Proc. 19th Pacific-Asia Conf. Knowledge Discovery and Data Mining, Springer, pp. 625-636 (2015)
7. LUCS-KDD. Software available at: http://cgi.csc.liv.ac.uk/ frans/KDD/Software/
8. Mahout. Software available at: http://mahout.apache.org/
9. Witten, I. H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2005)
10. Zida, S., Fournier-Viger, P., Lin, J. C.-W., Wu, C.-W., Tseng, V.S.: EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining. Proc. 14th Mexican Intern. Conf Artificial Intell., pp. 530-546 (2015)
11. Fournier-Viger, P., Lin, C.W., Duong, Q.-H., Dam, T.-L.: PHM: Mining Periodic High-Utility Itemsets . Proc. 16th Indust. Conf. Data Mining, 15 pages (2016)