Exploration of Textual Sequential Patterns

Hedi-Théo Sahraoui[1], Pierre Holat[2],
Peggy Cellier[3], Thierry Charnois[2], and Sébastien Ferré[1]

[1] Université de Rennes 1, IRISA, firstname.lastname@univ-rennes1.fr
[2] Université de Paris Nord, LIPN, firstname.lastname@lipn.univ-paris13.fr
[3] INSA Rennes, IRISA, Peggy.Cellier@irisa.fr
http://tal.lipn.univ-paris13.fr/sdmc/

1 Introduction

The extraction of regularities in texts is important for several natural language processing tasks. For instance, in information extraction, the regularities can allow to discover linguistic patterns [4] or to study the stylistics of authors [9]. When looking for those regularities, some specificities of textual data have to be taken into account: the sequentiality of the data (i.e., the order between words), the different levels of abstractions (i.e., words, lemma, Part-Of-Speech (POS) tags) and specific constraints (e.g., "the regularities have to contain a verb"). SDMC (Sequential Data Mining under Constraints) [3, 2][4] is a sequential pattern mining tool that deals with all those requirements. From a text, the tool extracts regularities called *sequential patterns*, i.e sequences of words, lemmas, and POS tags that frequently appear together in the text. In order to extract such patterns mixing different levels of abstraction, each word in the text is represented by itself but also by its lemma and its POS tags. In addition, SDMC allows to apply constraints to filter the extracted patterns: widespread constraints in data mining like minimum frequency (support) but also text-specific constraints like *"contains a verb"*.

A well-known drawback of pattern mining is the huge number of patterns that can be extracted. Even if SDMC manages the computation issue through constraints, the set of extracted patterns can be very large and hard to assess for users. In a previous work [5], the authors have used the Logical Information Systems (LIS) [6] paradigm to explore a set of patterns. The main advantage of this approach is that users can benefit from their background knowledge to navigate through the patterns.

In this paper we show how we have instantiated the LIS paradigm into SDMC to help users deal with their patterns and their texts. Indeed, we propose to explore the sequential patterns that appear in a text with a visualization of the sentences where those patterns occur. That exploration functionality is available online in the "Concordancier" menu as "Navigation dans les motifs". To illustrate the exploration, we use the French book "Le Petit Prince"[5].

[4] http://tal.lipn.univ-paris13.fr/sdmc/
[5] "Le Petit Prince", Antoine de Saint-Exupéry, 1943.

In: K. Bertet, D. Borchmann, P. Cellier, S. Ferré (Eds.): Supplementary Proceedings of ICFCA, Rennes, France, 2017.

Fig. 1. The textual interface of SDMC

2 Sequential Pattern Mining and SDMC

Sequential pattern mining [1] is a data mining technique that aims at discovering correlations between events through their order of appearance. It is an important field of data mining with broad applications (e.g. biology, marketing), and many algorithms to extract frequent sequential patterns [10, 8, 11]. In the context of the extraction of sequential patterns from texts, a *sequence* is an ordered list of distinct words also called *items*. Note that when considering different levels of abstraction for a word, a sequence is an ordered list of itemsets where each item represents some information about the word (e.g., the word itself, its lemma or a POS tag). The *support* of a sequence S in a text is the number of sentences in the text containing S. Given a mininimum support threshold *minsup*, the problem of frequent sequential pattern mining is to find the complete set of sequences whose support is greater or equal to *minsup*.

SDMC (Sequential Data Mining under Constraints) [3, 2] is an online sequential pattern mining tool with two user interfaces: one for mining textual data, and another for mining any kind of dataset. Figure 1 shows the first interface. SDMC handles various types of constraints which are not only numerical (e.g., the support constraint) but also symbolic and syntactic (e.g., "the pattern has to contain a verb"). These multiple constraints enable the user to express a large scope of knowledge to focus on interesting textual sequential patterns.

3 Logical Information Systems (LIS)

Logical Information Systems (LIS) [6] are a paradigm of information retrieval that combines querying and navigation. LIS are formally based on Logical Con-

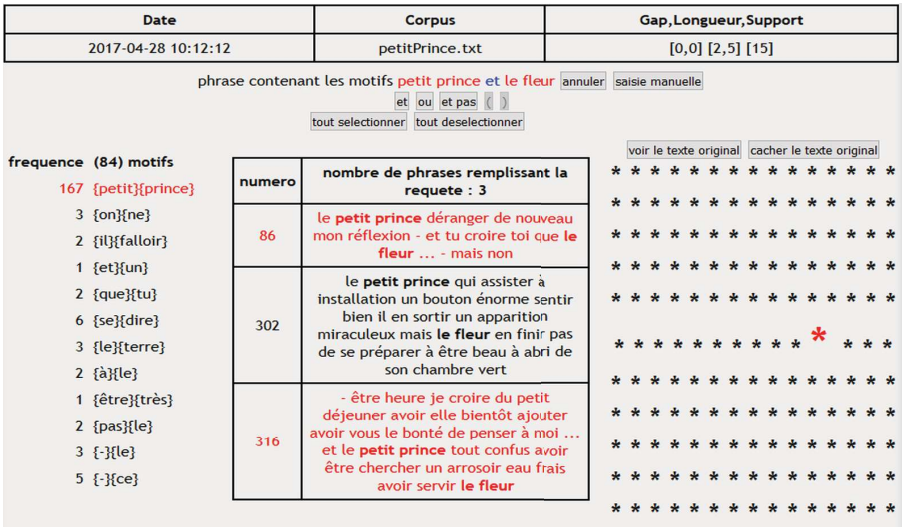

Fig. 2. The LIS interface of SDMC to explore a text and its patterns

cept Analysis (LCA), a logical generalization of Formal Concept Analysis [7]. In LCA, objects are described by logical formulas rather than sets of attributes. The concept lattice serves as a navigation structure where each concept is a navigation place. Because of its huge size, LIS only show a local view of the concept lattice, centered on each navigation place. A local view has three components: (a) the query that is a Boolean combination of descriptors, (b) the extent that is the set of objects matching the query, and (c) the index that is the set of descriptors that occur in the extent, along with their relative frequency. The index descriptors can be used as navigation links to modify the query, and hence reach related concepts: e.g., adding a descriptor to reach a more specific concept.

4 LIS Exploration Interface in SDMC

We have instantiated the LIS paradigm to textual sequential patterns by considering sentences as objects, and sequential patterns as descriptors. This has been implemented and integrated into SDMC as a new user interface. Figure 2 shows that LIS interface in SDMC.

On the top of the screen, information about the extraction are given: the date of the extraction, the corpus (text), and the values of numerical constraints used for pattern extraction: [min,max] gap size between words in patterns, [min,max] length of the extracted patterns, and minimum support. The query appears just below those information. It is a Boolean combination of patterns. Here, it is a conjunction of two patterns: "phrases contenant les motifs **petit prince** et **le fleur**" (in English "sentences that contain patterns **little prince** and **the**

flower") which means that the user has selected pattern "**petit** followed by **prince**" and pattern "**le** followed by **fleur**".

On the main part of the screen, there are 3 parts. On the left part, there is the index, i.e. textual patterns that can be selected or added to the query. On the middle part, there is the extent, i.e. the sentences that match the query. In the example, three sentences (82, 302, and 316) contain both patterns **petit prince** and **le fleur**. On the right part, a text view is displayed where the selected sentences appear in bold and red. The tool proposes three kinds of text views: the text itself, a compact version where the sentences are replaced by stars (as shown on the figure), and a void view (useful for very long texts).

5 Conclusion

In this paper we have presented an interface to explore regularities extracted from text, called *textual sequential patterns*. The exploration is based on a conceptual navigation over the set of all patterns in the logical information systems framework, a logical version of formal concept analysis. The exploration interface is available through the online tool SDMC.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE. pp. 3–14. IEEE Computer Society (1995)
2. Béchet, N., Cellier, P., Charnois, T., Crémilleux, B.: SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes (2013)
3. Béchet, N., Cellier, P., Charnois, T., Crémilleux, B.: Sequence mining under multiple constraints. In: Wainwright, R.L., Corchado, J.M., Bechini, A., Hong, J. (eds.) ACM Symposium on Applied Computing. pp. 908–914. ACM (2015)
4. Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Crémilleux, B., Gandrillon, O., Kléma, J., Manguin, J.: Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. J. Biomedical Semantics 6, 27 (2015)
5. Cellier, P., Ferré, S., Ducassé, M., Charnois, T.: Partial Orders and Logical Concept Analysis to Explore Patterns Extracted by Data Mining. In: Conceptual Structures for Discovering Knowledge. pp. 77–90. Springer, Berlin, Heidelberg (Jul 2011)
6. Ferré, S., Ridoux, O.: Introduction to logical information systems. Inf. Process. Manage. 40(3), 383–419 (Jan 2004)
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag New York, Inc. (1997)
8. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H.: PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: ICDE (2001)
9. Quiniou, S., Cellier, P., Charnois, T., Legallois, D.: What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In: Int. Conf. on Computational Linguistics and Intelligent Text Processing. LNCS, Springer (2012)
10. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: EDBT (1996)
11. Zaki, M.: Spade: An efficient algorithm for mining frequent sequences. Machine Learning 42(1/2), 31–60 (2001)