Sequential Pattern Mining of Multimodal Streams in the Humanities

Marwan Hassani [•] Christian Beecks [•] Daniel Töws [•]
Tatiana Serbina [○] Max Haberstroh [§] Paula Niemietz [○]
Sabina Jeschke [§] Stella Neumann [○] Thomas Seidl [•]
[•]Data Management and Data Exploration Group
{hassani, beecks, toews, seidl}@cs.rwth-aachen.de
[○]Group of English Studies [§]IMA/ZLW & IfU
RWTH Aachen University, Germany

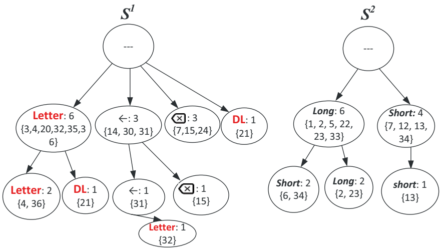
Abstract: Research in the humanities is increasingly attracted by data mining and data management techniques in order to efficiently deal with complex scientific corpora. Particularly, the exploration of hidden patterns within different types of data streams arising from psycholinguistic experiments is of growing interest in the area of translation process research. In order to support psycholinguistic experts in quantitatively discovering the non-self-explanatory behavior of the data, we propose the *e-cosmos miner* framework for mining, generating and visualizing sequential patterns hidden within multimodal streaming data. The introduced $\mathcal{M}$SS-BE algorithm, based on the *PrefixSpan* method, searches for sequential patterns within multiple streaming inputs arriving from eye tracking and keystroke logging data recorded during translation tasks. The *e-cosmos miner* enables psycholinguistic experts to select different sequential patterns as they appear in the translation process, compare the evolving changes of their statistics during the process and track their occurrences within a special simulator.

1 Introduction

In the area of translation process research, psycholinguistic experiments are conducted to draw conclusions on the cognitive processing that takes place during translation tasks. The cognitive dimension, operationalized in terms of gaze- and keystroke-related statistics, can be used to explain, among other things, linguistic features of the translated language identified with the help of corpus studies (analysis of large and representative collections of translated and non-translated texts). An example of a research question is the analysis of grammatical complexity both in the final translation product and in the intermediate versions of the translation process: the psycholinguistic perspective can extend the corpus-based analysis of the phenomenon by answering such questions as whether grammatically more complex items require more cognitive processing and what translation strategies are employed to produce a more or less complex linguistic structure [APN^{+}10]. One of the aims of the *e-cosmos* project is to treat the process data, i.e. the unfolding translations, as another type of corpus . This corpus will be queried for both behavioural and linguistic information, which researchers can analyze quantitatively while still taking into account fine-grained details. The application of data-mining techniques allows us to investigate the data in a bottom-up manner; this approach yields multimodal patterns which may otherwise be overlooked yet contribute to the explanatory potential of the data. Within the *e-cosmos* project, we are asked to analyze the multimodal data collected during a psycholinguistic translation experiment. Sixteen participants (eight professional translators and eight domain specialists, i.e. PhD students of physics) were asked to translate an abridged popular-scientific text from English into their native language (German). Their keystrokes, mouse movements and pauses in between were recorded . Additionally, eye-tracking data during this translation task was gathered using the remote eye tracker *Tobii* 2150. It contains information on individual gaze points that have been collected with a 50 Hz sampling rate (50 gaze points per second) for each translation session. The remainder of this paper is organized as follows: Section 2 introduces the $\mathcal{M}$SS-BE algorithm used mainly by our demo, then in Section 3, we explain the analysis and interpretation of patterns using the proposed *e-cosmos miner* demonstrator as well as the demo plan.

2 The $\mathcal{M}$SS-BE Algorithm

The SS-BE algorithm was suggested in [MDH08] to find sequential patterns in a single data stream. This algorithm breaks a stream into batches and performs a single scan over every batch using the *PrefixSpan* algorithm [PHMA^{+}04]. It promises a bounded error by guaranteeing that all true patterns in any batch are output at the end of that batch. In [HS11], some ideas were discussed on using the SS-BE algorithm for mining multiple streams with the aim of gaining a health context prediction. The SS-BE algorithm has a nice, stream-friendly, feature of pruning and updating the $T0$ lexicographic tree containing the PrefixSpan output. However, it has two main problems which make it insufficient for dealing with the multimodal streams in our humanities scenario: (1) It is unable to discover patterns within multiple streams, (2) Its batch-based method might results in losing all interesting sequential patterns whose sub-sequences appear on two batches. We propose in this demo the $\mathcal{M}$SS-BE algorithm (Multiple Stream Sequence miner using Bounded Error) in order to overcome the aforementioned drawbacks. $\mathcal{M}$SS-BE seeks sequential patterns that consist of consecutive patterns belonging to multiple streams The multiple stream mining concept is mainly based on the a-priori observation. When using an equal or larger threshold value, any sequence of consecutive patterns that belong to multiple streams is *frequent* if and only if these consecutive patterns were *already frequent*. The $\mathcal{M}$SS-BE algorithm we suggest in this demo can be summarized in the following main steps: **1.** Generate all sequential patterns appearing in every stream for finding *intra-stream* sequential patterns. Input are patterns that consist of single-item, timestamped sequences. These timestamps for each batch are added to the end of that sequence in $\mathcal{M}$*T0* tree structure (a modified structure of [MDH08]). An example of the output is detailed in Fig. 1 (Top), by assuming that a *window width* = 40 and *min_supp*= $\frac{1}{40}$ (i.e. a sequence is frequent if it appears at least once). Paths consisting of more than one node are representing *consecutive* sequences frequently seen. **2.** To find the *inter-stream* sequential patterns, for every window, a table similar to that in Fig. 1 (Bottom) is built. Using a *min_supp*= $\frac{2}{40}$ on the inter-stream level, the resulted inter-stream sequential pattern from our example is: $\langle LE, lo\rangle$:2, which represents a two-times appearance of typing a letter

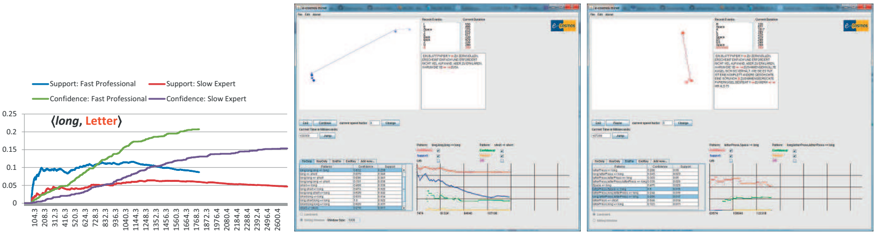


t_i	1	2	3	4	5	6	7	12	13	14	15	20	21	22	23	24	30	31	32	33	34	35	36
S^1	...	...	LE	LE	...	...	⌫	...	...	←	⌫	LE	DL	...	...	⌫	←	←	LE	...	...	LE	LE
S^2	lo	lo	...	...	lo	sh	...	sh	sh	...	...	...	...	lo	lo	...	...	...	...	lo	sh	...	...

Figure 1: **Top:** An example of the state of the $\mathcal{M}T0$ tree for streams S^1: *KeyStrokes* and S^2: *Eye Gazes* showing in each node: [*sequential pattern*: *count* $\{$*ending timestamps*$\}$]. **Bottom:** The timing sequence of sequential patterns for generating the *inter-stream* rules based on consecutive sequential patterns that belong to different streams. *LE= letter, DL= delete, sh= short gaze, lo= long gaze*.

followed by a long gaze in the current window. **3.** Continuously maintain the nodes in the $\mathcal{M}T0$ tree as the window slides (no batch processing). Delete an item if all its ending timestamps are older than the current starting timestamp of the window. Decrement its count if one of its ending timestamps older than the window starting timestamp. Insert newly seen items in their correct available/new place in the tree and increment/initiate the count. Delete the nodes that represent infrequent items.

3 Analysis of Patterns With the *e-cosmos Miner* Demo



(a) (b) (c)

Figure 2: (a) A Comparison between a fast professional translator and a slow domain expert (physicist) translator over the time (in Seconds) obtained by the *e-cosmos miner* for the pattern: $\langle long, Letter\rangle$. Screen shots of the *e-cosmos miner* comparing: (a) Two intra-stream rules with the sliding window concept, (b) Two multimodal rules with the landmark window concept.

The *e-cosmos miner* reads the eye tracking and keystroke logging in a streaming way to observe the changes of the patterns as the translation session evolves. The *e-cosmos miner* currently identifies *intra-stream* (single) and *inter-stream* (multimodal) behavioral sequential patterns consisting of the following elements: short or long fixations as well as different types of characters typed by the participants (cf. Fig. 1). Previous research has established that average fixation duration during silent reading (for comprehension) is approximately 200 milliseconds [JJ08]. A distinction for the characters was done to differentiate normal from other control characters as: Delete, Backspace, Arrows, left mouse click, etc. (cf. Fig. 1).

An example of an identified multimodal pattern is a long fixation followed by a letter.

Since long fixations are typically interpreted as one of the concrete indicators for increased cognitive demand , this pattern might appear during processing of the linguistic unit that is produced after a long fixation. In Fig. 2(a), we picked a professional translator who required the least amount of time for translation to compare them to the domain specialist characterized by the longest translation session. Some statistics of the rule: $long \Rightarrow Letter$ which is implied by the pattern $\langle long, Letter\rangle$ are depicted in Fig. 2(a). The changes of the confidence and the the support values over time are different for the two participants. The figure shows a first indication of potential differences between the two groups. Moreover, these first patterns contribute both to an understanding of how to appraise patterns in general, and to increasing sophistication in the types of patterns suggested by the *e-cosmos miner*. As can be seen in Fig. 2, an external file containing the timestamped gazes and keystrokes can be read within the *e-cosmos miner*. By selecting the size of the sliding window, or selecting the landmark window, clicking the *start* button (which will flip to be an *end* button) will start the simulation. The recent eye gazes, their duration, their location and movements are simulated in the upper part of the GUI together with the current key strokes and the resulting text. All emerging patterns are depicted in the bottom and the statistics like: (confidence, support and lift) are plotted against the evolution of the real time in the bottom right corner. The psycholinguistic expert can filter the observed patterns according to: one-dimensional vs. multimodal (cf. Fig. 2(b) vs. Fig. 2(c)), or according to stream of the ending item of the pattern. Additionally, the simulation and generation of patterns can jump to any time of the session, and the simulation speed can be increased or decreased. BTW participants can observe the effect of varying different streaming parameters over the process of building the $\mathcal{M}T0$ tree and the intermediate top-k patterns of a single and of multiple translators and match those to the current translation step. The *e-cosmos miner* can be downloaded from our website: `http://dme.rwth-aachen.de/en/research/projects/e-cosmos/ecosmos_miner`.

Acknowledgments

Funded by the Excellence Initiative of the German federal and state governments.

References

[APN$^+$10] F. Alves, A. Pagano, S. Neumann, E. Steiner, and S. Hansen-Schirra. Units of translation and grammatical shifts: towards an integration of product- and process-based research in translation. In *Translation and Cognition*, pages 109 –142. 2010.

[HS11] Marwan Hassani and Thomas Seidl. Towards a Mobile Health Context Prediction: Sequential Pattern Mining in Multiple Streams. In *MDM '11*, pages 55–57, 2011.

[JJ08] A. L. Jakobsen and K.T.H Jensen. Eye movement behaviour across four different types of reading task. In *Looking at eyes: eye-tracking studies of reading and translation processing*, pages 103 –124. 2008.

[MDH08] L.F. Mendes, Bolin Ding, and Jiawei Han. Stream Sequential Pattern Mining with Precise Error Bounds. In *ICDM '08*, pages 941 –946, 2008.

[PHMA$^+$04] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. *IEEE Trans. on KDE*, 16:1424–1440, 2004.