Visualisation Analysis for Exploring Prerequisite Relations in Textbooks

Samuele Passalacqua, Frosina Koceva, Chiara Alzetta
Ilaria Torre, and Giovanni Adorni

University of Genoa, Italy, Department of Informatics, Bioengineering, Robotics and Systems Engineering
samuele.passalacqua@dibris.unige.it
{frosina.koceva,chiara.alzetta}@edu.unige.it
{ilaria.torre,giovanni.adorni}@unige.it

Abstract. Building automatic strategies for organising knowledge contained in textbooks has a tremendous potential to enhance meaningful learning. Automatic identification of prerequisite relation (PR) between concepts in a textbook is a well-known way for knowledge structuring, yet it is still an open issue. Our research contributes for better understanding and exploring the phenomenon of PR in textbooks, by providing a collection of visualisation techniques for PR exploration and analysis, that we used for the design of and then the refinement of our algorithm for PR extraction.

Keywords: prerequisite relation · knowledge structuring · information visualisation.

1 Introduction and Background

In our age we are experiencing an increasing availability of learning resources and self-regulated learning. In this scenario, the development of automatic strategies for structuring knowledge is motivated by the need of curricula planning. In particular, organising the knowledge contained in a textbook and structuring it as a knowledge map that makes an explicit representation of prerequisite dependencies between concepts has a formidable potential for building intelligent content, authoring systems for instructional design and e-learning applications [9, 8, 19, 16]. However, the manual construction of structured knowledge from teaching materials requires an additional and substantial workload provided by experts. Consequently, the research presented in this paper pursues the extraction of prerequisite relation [5, 10, 20, 14, 1] (**PR**, henceforth) and investigates the use of information visualisation techniques for better understanding and exploring this phenomenon and its characteristics in textbooks.

The PR relation is a dependency relation defining precedence between two concepts t_u and t_v: it represents what a learner must know/study (concept t_u) before approaching concept t_v, where a concept can be seen as an atomic piece of knowledge of the subject domain. By definition, the main properties of a PR

relation are the followings: (1) binary relation: it involves pairs of concepts; (2) anti-reflexive relation: concept t_u cannot be a prerequisite of itself; (3) transitive relation: if $t_u \prec t_v$ and $t_v \prec t_z$, than $t_u \prec t_z$. As a result of these conditions, the key concepts of the textbook can be represented as nodes in a directed acyclic graph $\mathcal{G}$ related to each other by means of PR relations.

The effective integration of visualisation technologies in curricula with the purpose of facilitating teaching and learning of abstract concepts has been already investigated (see for instance [17] for a study on visualisation and learner's engagement in Computer Science education). More recently, in Educational Data Mining (see [18] for a survey), several studies are oriented toward visualising different kinds of educational data. In the field of Learning Analytics, information visualisation techniques have been studied to empower learning dashboards with graphical representations of the learning process [7]. More in general, Visual Analytics aims to handle large amounts of multidimensional data by means of interactive graphic interfaces and advanced visual representation techniques during the process of analysis [11]. To the best of our knowledge, a specific contribution on how information visualisation techniques can be applied to the analysis of prerequisite relations in textbooks is still missing in the literature. [15] employed two representations (Hierarchical Edge Bundling and Hive Plots) on the structure of a book to show how these visualisations can deal with large graphs that have a hierarchical nature. However, he did not further develop the investigation on textbook prerequisites by means of visualisation.

Our research on PR extraction from textbooks is enhanced by the use of Information Visualisation techniques in the following phases:

(i) Exploring and discovering insights of PR;
(ii) Refining the algorithm of PR extraction by means of visual analysis of patterns and comparison between gold standard PR graphs vs extracted PR graph.

In the first phase (*i*), visualisation analysis techniques were applied to a concept map manually created by experts. The purpose of map creation was to make explicit the pedagogical relations among concepts in the textbook, while the aim of visualisation analysis was to discover new insights into PR. The dataset was explored through matrix and graph visualisations, both enhanced with filtering and ordering functions. This analysis supported the definition of the algorithm in [1] for PR extraction.

In the second phase (*ii*), visualisation analysis was applied on a map automatically extracted from a textbook using the strategy described in [1]. We applied visualisation analysis with the aim of improving pattern discovery, refining the algorithm and better understanding how the automatic approach is affected by changing the parameters. In this phase we relied on a gantt representation of the algorithm results. Further analysis was conducted by "visually" comparing the extracted map and the gold map with the purpose of analysing graph differences at various levels.

Most of the information visualisation analysis tools that we propose are meant for the analyst (e.g., researcher) who intends to discover new insights

or confirm existing hypotheses on the PR. Nevertheless, some of these techniques/tools give a graphical visualisation that can be potentially useful also for learners, teachers or instructional designers [22, 7, 6]. For example, from this perspective a graph representation can be proposed as a supporting tool in a question answering scenario where the underneath knowledge structure is used to retrieve the most appropriate learning path without leaving out prerequisite concepts [2]. Such a tool can produce a graphical representation that reflects and explicates the necessary prerequisite knowledge or deepening knowledge in respect of the learner's query. While the latter user and teacher-centric case is left for future works, in the rest of this paper we will focus on (*i*) and (*ii*).

In the following we describe our approach, techniques and data used for visualisation analysis for both the phases described above (i.e., PR exploration in Section 2 and PR extraction and algorithm refinement in Section 3), and for each phase we discuss the results. The visualisation analysis tools proposed in this paper are available at (teldh.dibris.unige.it/projects/).

2 PR exploration

Gold Dataset Five experts were asked to read a network related chapter from a Computer Science textbook [4] and annotate the prerequisite concepts of each relevant term appearing in the text. Each annotator was provided with the same initial set of concepts extracted with the semi-automatic strategy described in [1]. Besides these terms, each expert could independently add new concepts to the terminology if they regard them as relevant. Experts produced different sets of concept pairs annotated with PRs and the final gold dataset resulted from the combination of each expert annotation. To achieve this goal, all pairs of concepts annotated by at least one expert were added to the dataset as positive examples (i.e. showing a PR). Negative examples (i.e. non-PR concept pairs) were automatically created by pairing all concepts. Only those pairs that were not annotated by any expert were added as negative pairs. The final output is a binary-labelled dataset presenting 124.609 concept pairs in total, obtained by all possible combinations of 353 concepts. The dataset is really sparse since, among all relations, only 1052 show a PR (0.84%).

Concept Graphs. Several variants of network-like representations (see for instance **Fig. 1**) have been used during PR exploration to visually detect elements such as loops (as resulting from human errors during the process of annotation) and transitive edges. However, as the dataset becomes larger, a concept graph becomes harder to explore, especially if no filtering functions are implemented. In this case, other forms of visualisation are more effective.

Concept Matrix Chart. This is a dynamic and interactive representation of a $|T| \times |T|$ asymmetric adjacency matrix M, where each colored cell $M_{i,j}$ represents a prerequisite relation between concepts i and j (see **Fig. 2**). Different colors can help to visually differentiate clusters of concepts, as they have been recognised by a community detection algorithm [1]. Intuitively these clusters shows

[1] In our implementation depicted in **Fig.2** we used the Infomap algorithm.

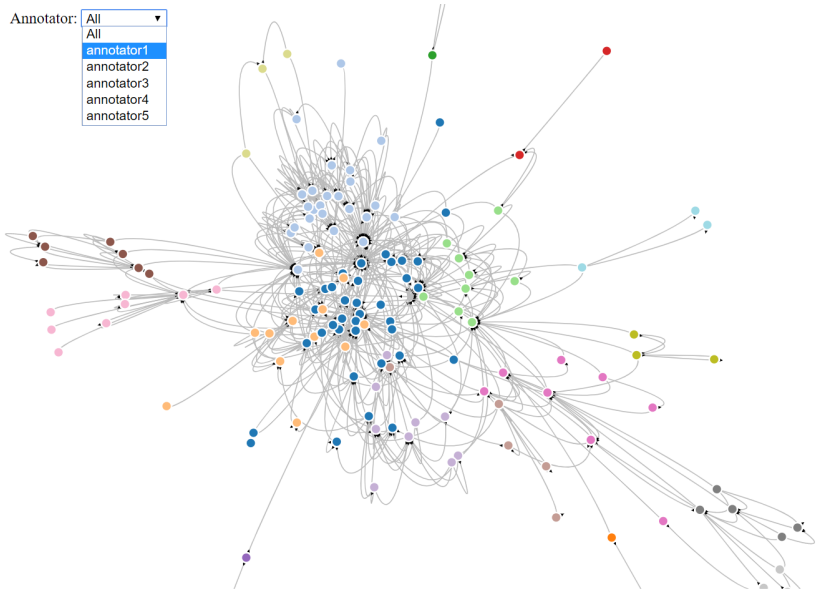

Fig. 1. A Concept Graph that allows decomposition in sub-graphs belonging to individual annotators.

the membership of a concept within a thematic unit (e.g., concepts related to network security, or to network classification, and so on). Different shades of the same color can be used to encode different degrees of inter-agreement among annotators (if M is used to visually depict a gold standard) or different scores (if M represents the output of an automatic method). The matrix arrangement is dynamic, i.e. the concepts along the matrix can be sorted according to different criteria: order of first appearance in the text, alphabetical order, frequency and cluster membership.

Discussions and Results. The analysis performed on the concept graph built from the gold standard allowed to reveal interesting properties concerning graph's transitivity, topology and connectivity. By comparing subgraphs belonging to different experts, we discovered that the number of transitive edges largely varies from annotator to annotator. Thanks to this observation, we discussed the phenomenon with the annotators, ascertaining that their choices depend on a different interpretation given to the meaning of a distant or weak prerequisite relation. Some of the experts tend to think in terms of graph paths, while others in terms of didactic sequences. As an example, the relation between "computer" and "local area network" (LAN) can be seen on the one hand as a transitive relation (if one has in mind the path in the graph connecting the first concept to the second by means of several bridging concepts in the middle), but on the other hand it can also be seen as a direct prerequisite relation (if one realises that

"computer" is a fundamental notion, without which a student cannot possibly hope to understand what a LAN is). Concerning the topology, graph visualisation confirmed our intuition that prerequisite relations do not necessarily replicate ontological relations. As an example, let us take a pair of concepts such as "client side" and "server side": in a domain ontology these would very probably be represented as sibling nodes, but we cannot always expect the same behaviour when approaching a didactic text. In similar contexts, even if a co-requisite relation would seem the most natural choice of presenting these kind of concepts (i.e., the author explains them together, and the former is not a prerequisite of the latter, nor vice versa), a prerequisite relation is still possible (e.g., if the author first explains the former and then relies on the knowledge gained by the reader to explain the latter). As can be noted in the graph, hypernym-hyponym and holonym-meronym relations deserve a similar discussion. External lexical resources would typically categorise pairs of words such as "device" (broader, hence at top-level) and "hub" (narrower, hence at bottom-level) or "byte" (the whole) and "bit" (the part of) in a hierarchical manner. Conversely, in textbooks (sometimes even in the same textbook) we can easily find both top-down and bottom-up explanations. Lastly, the connectivity of the graph (which we discussed above as influenced by the annotators' perception of what a prerequisite relation is) also largely depends on the annotator's level of domain knowledge.

The analysis performed on the Concept Matrix Chart built from the gold dataset revealed an important insight for the direction of the prerequisite relation. After applying the first sorting criterion (i.e., order of first appearance), the matrix tends towards an upper triangular, with colored cells mostly concentrated in the area that is slightly above the diagonal. This pattern confirms the hypothesis that prerequisite relation is highly correlated with co-occurrence and temporal order . Consequently, the temporal order of concepts is a reliable criterion to assign a direction to relations that are automatically extracted by an algorithm. The most notable exception in this pattern is represented by concepts such as "computer" or "network", which tend to be spread across the entire row of the matrix. However, this phenomenon is due to the fact that these are the main concepts of the whole chapter of the textbook, hence they frequently reoccur along the entire text and moreover they could commonly be prerequisites (rather than subsidiaries) of many other concepts.

The analyses above supported the definition of the algorithm presented in [1] for PR extraction, which is based on Burst analysis and temporal order.

3 Algorithm Refinement

Burst Dataset The method devised to obtain the Burst Map dataset exploits burst analysis [12] based on co-occurrence of relevant terms in a text and combined with temporal ordering, as described in [1]. Burst analysis is based on the observation of *burst intervals* of a phenomenon, that is the periods of time when the phenomenon is particularly relevant along a time series (i.e., its occurrence rises above a certain threshold) [12]. Following [21, 13], we applied burst analysis

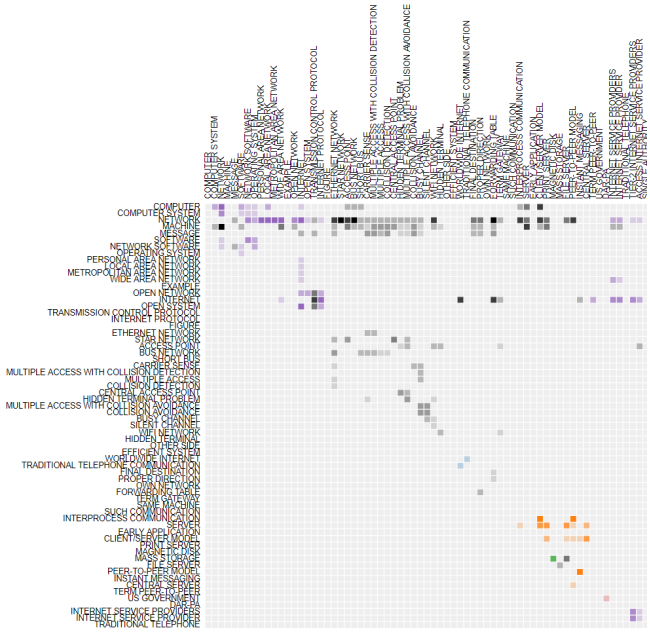

Fig. 2. Concept Matrix Chart

to detect the bursting intervals of relevant terms along the textbook chapter and analyse different types of temporal patterns established by two concepts by applying spatial-temporal reasoning on the extracted patterns in order to identify PR relations. To capture and formalise their temporal relations, we exploited a subset of temporal relation defined by Allen's interval algebra [3]. Our selection is shown **Fig. 3**. The result of the process is a concept graph with 353 concept nodes and 124,256 possible pairs of distinct concepts related by a PR.

Burst Gantt Chart. This is a Gantt diagram showing bursts of concepts along the horizontal temporal axis (time can be measured in sentences or tokens), while concepts are arranged along the vertical axis, according to their temporal order (see **Fig. 4**). The main purpose of this visualisation is facilitating the analysis of temporal patterns between intervals of different concepts.

Moreover, as the chart incorporates data taken from three different sources (the output of the burst algorithm, the gold dataset and the textbook itself), we can use it to perform further kinds of analysis and textbook exploration. For instance, by clicking on a concept label in the vertical axis, we can compare Allen's temporal relations and gold relations and thus investigate possible matches (see **Fig. 5**).

By clicking on a burst we can instead read the portion of the textbook covered by that interval (see **Fig. 6**). This procedure enables us to easily find blocks of sentences where a concept is introduced for the first time, then resumed (with

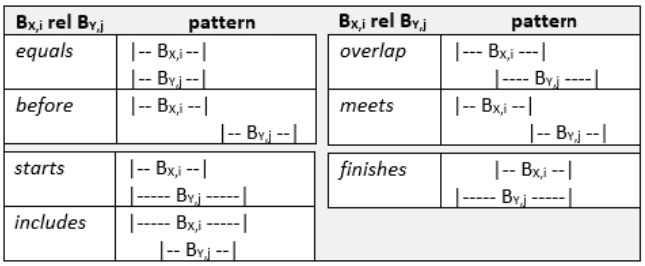

Fig. 3. Allen's patterns that map PR relations betwen Burst of concepts X and Y.

or without another concept) and eventually left behind. Vertical partition lines have been drawn to indicate boundaries between sections, while other sorts of markers can be traced near the temporal axis to identify sequences of sentences that according to experts are particularly rich of prerequisite relations.

Concept Graph with Allen's Relations. For investigating Allen's temporal patterns and prerequisite relations, we also propose to transform the Burst Gantt Chart into a weighted directed and edge-labeled graph G_A, where edges are labelled using Allen's algebra. For each two distinct concepts X and Y in the Burst Gantt Chart, if a pair of bursts $B_{x,i}$ and $B_{y,j}$ is related n times by Allen's relation a, we represent this configuration in G_A as $X \rightarrow_{(a,n)} Y$, where (a, n) are the edge label and edge weight respectively.

In this representation, bursts are collapsed into one node for each concept, while multiple edges are maintained and a weight is assigned to them according to how many times that temporal relation occurs between bursts of those two concepts. The aim of this conversion is producing a graph that can be compared with the gold standard graph, as a means for achieving more confidence on the weights that should be assigned to the different Allen's patterns. The result of the conversion is a highly connected graph which needs to be explored with filters, e.g., filtering concepts that have different relevance, or filtering by specific Allen's relation or combination of relations, as well as filtering according to edges or nodes weights. As displayed in **Fig. 7**, different colors for edges show different Allen's relations, while the width is proportional to the number of times an Allen's relation is founded between two concepts. The dimension of a node is proportional to the importance of the concept (this value can be measured using frequency, relevance or summing all the lengths of its bursting periods in the text). In our implementation we also used different colors to encode concepts that in the gold standard are sources, sinks or internal nodes. In the first case the node has zero indegree and this means that for annotators it represents a primary notion—a concept already known by the learner; in the second case the node has zero outdegree and thus it may be intended as a final learning outcome.

Discussions and Results. Allen's patterns allow to capture PR relations quite well [1], however they overestimate the PR relations. This comes as a

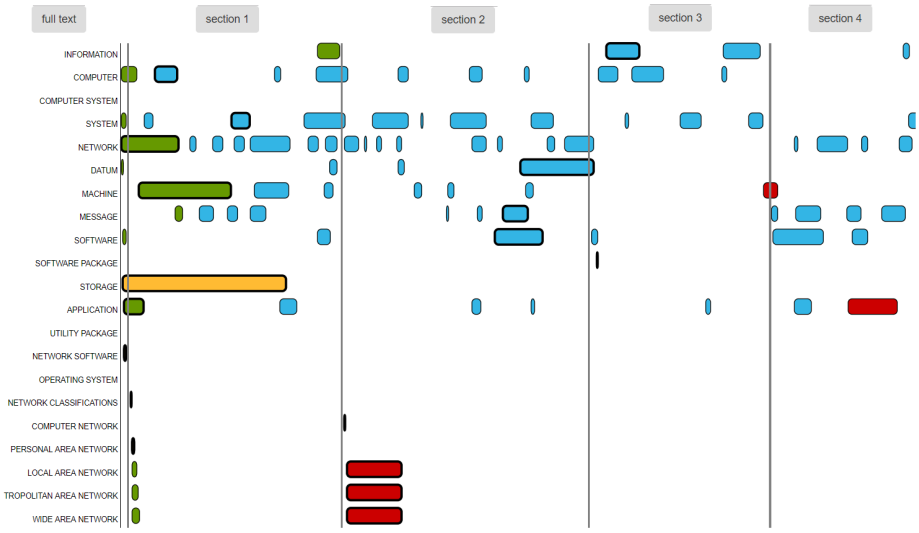

Fig. 4. Burst Gantt Chart

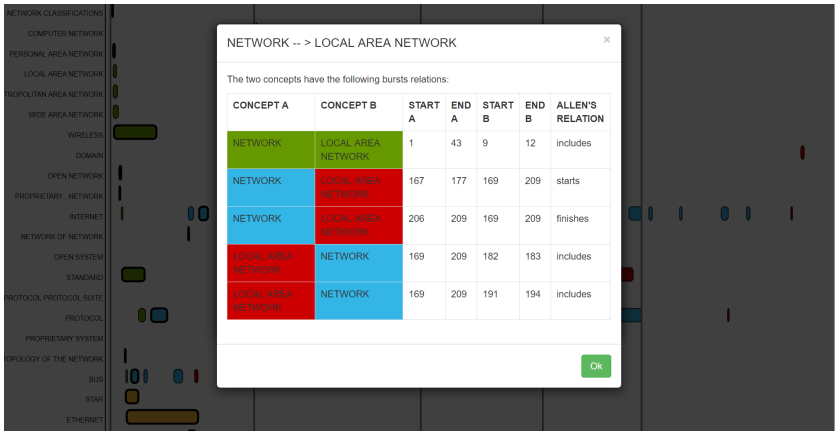

Fig. 5. Analysis of temporal patterns.

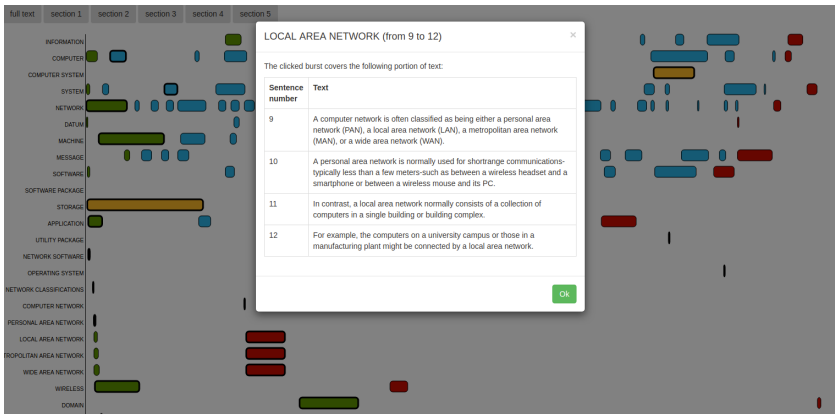

Fig. 6. Textbook Exploration with Gantt

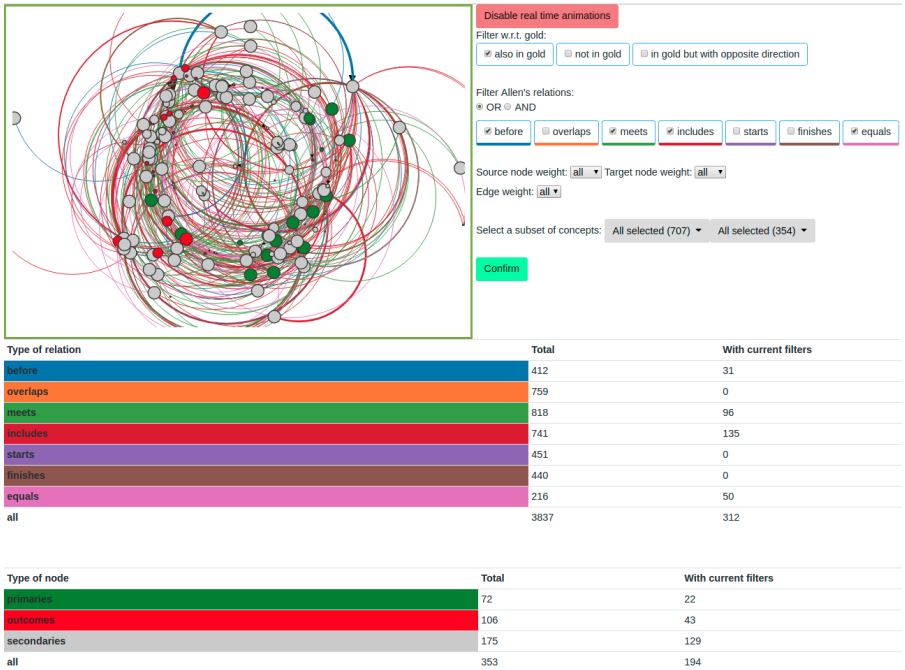

Fig. 7. Concept Graph with Allen's Relations

straightforward observation considering the Concept Allen Graph visualisation. As it can be seen, the number of detected Allen's relations is much bigger than the set of relations identified by the experts (even when transitive closure is applied on the experts' graph in order to reduce variety in the number of transitive edges).

Therefore, the Burst Gantt Chart was used to analyse possible combinations of Allen's patterns and more sophisticated conditions that should be satisfied between bursts of two concepts. As a result of the aforementioned analyses, we observed that Allen's Algebra, as used in our Burst-based algorithm, is likely to fail when an Allen relation is identified between bursts of concepts X and Y, but no bursts of X are present in the text before that relation. This is consistent with the intuitive consideration that concept X should be introduced before Y in order to be a prerequisite of Y, and thus for two concepts X and Y, a necessary but not sufficient condition in order to have X prerequisite of Y is that X should be previously explained, i.e., $|B_X| > 1$. Considering bursts instead of simple occurrences of a term allows to exclude cases where X occurs before Y and X is not really explained but rather simply introduced (the analysis of the text showed for example several cases where the content of the next section is mentioned before, as a guide for the reader).

As future work, we plan to implement refinements of the algorithm that take this into account. This is not trivial, since for instance the condition $|B_X| > 1$ does not apply in cases where X is a primary notion, namely a concept already known by the learner as background knowledge.

Furthermore, we plan to use Concept Allen Graph to explore combinations of Allen's patterns by filtering them in conjunction or disjunction and comparing the results with the gold standard.

4 Conclusion

In this paper we presented a collection of visualisation techniques conceived to help researchers and analysts in their effort of better understanding the issue of prerequisite dependencies in textbooks and developing more powerful strategies for the automatic extraction, with the final aim of giving a contribution to the field of intelligent textbooks. The results of our analysis support the hypotheses regarding the correlation between PR direction and temporal concept ordering. Furthermore, visual analysis of the PR algorithm provides valid insights on the burst patterns combination. The tools for PR exploration presented in this paper are available online[2] as a support for the community of researchers working on the analysis of prerequisite relations.

Our future work includes enhancing these techniques with further functionalities and applying them, in conjunction with new techniques as well, in different contexts of use and a larger variety of educational texts. New functionalities can be implemented in order to broaden the scope of practice allowed by the visualisation tools. For instance, the Gantt Chart could also be used as an instrument

[2] Prerequisite Extraction from TextBooks at http://teldh.dibris.unige.it/projects/

for doing validation, i.e. thanks to it the expert can validate the individual patterns revealed by the algorithm. The Matrix Chart can be used not only for exploring PR in the experts' annotation (i.e. phase (i)), but also for algorithm refinement (phase (ii)), for example in cases of co-occurrence and/or temporal based algorithms. Finally, we are also working on techniques that more directly address the needs of learners and teachers in their common activities of selecting, accessing, exploring and organising learning materials.

References

1. G. Adorni, C. Alzetta, F. Koceva, S. Passalacqua, and I. Torre. Towards the identification of propaedeutic relations in textbooks. In *International Conference on Artificial Intelligence in Education*. Springer, 2019.
2. G. Adorni and F. Koceva. Educational concept maps for personalized learning path generation. In *Conference of the Italian Association for Artificial Intelligence*, pages 135–148. Springer, 2016.
3. J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 1983.
4. G. Brookshear and D. Brylow. *Computer Science: An Overview, Global Edition*, chapter 4 Networking and the Internet. Pearson Education Limited., 2015.
5. D. S. Chaplot, Y. Yang, J. G. Carbonell, and K. R. Koedinger. Data-driven automated induction of prerequisite structure graphs. In *EDM*, pages 318–323, 2016.
6. C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. Visualizing patterns of student engagement and performance in moocs. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 83–92. ACM, 2014.
7. E. Duval. Attention please!: learning analytics for visualization and recommendation. *LAK*, 11:9–17, 2011.
8. J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, and X. Hu. *Knowledge spaces: Applications in education*. Springer Science & Business Media, 2013.
9. J.-C. Falmagne and J.-P. Doignon. *Learning spaces: Interdisciplinary applied mathematics*. Springer Science & Business Media, 2010.
10. J. Gordon, L. Zhu, A. Galstyan, P. Natarajan, and G. Burns. Modeling concept dependencies in a scientific corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 866–875, 2016.
11. D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Tenth International Conference on Information Visualisation (IV'06)*, pages 9–16. IEEE, 2006.
12. J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
13. S. Lee, Y. Park, and W. C. Yoon. Burst analysis for automatic concept map creation with a single document. *Expert Systems with Applications*, 42(22):8817–8829, 2015.
14. C. Liang, J. Ye, S. Wang, B. Pursel, and C. L. Giles. Investigating active learning for concept prerequisite learning. *Proc. EAAI*, 2018.
15. T. S. McTavish. Facilitating graph interpretation via interactive hierarchical edges. In *EDM (Workshops)*, 2014.

16. R. Mizoguchi and J. Bourdeau. Using ontological engineering to overcome ai-ed problems: Contribution, impact and perspectives. *International Journal of Artificial Intelligence in Education*, 26(1):91–106, Mar 2016.
17. T. L. Naps, G. Rößling, V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger, et al. Exploring the role of visualization and engagement in computer science education. In *ACM Sigcse Bulletin*, pages 131–152. ACM, 2002.
18. C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.
19. R. J. Shavelson. Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, 11(3):231–249, 1974.
20. S. Wang, A. Ororbia, Z. Wu, K. Williams, C. Liang, B. Pursel, and C. L. Giles. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the 25th acm international on conference on information and knowledge management*, pages 317–326. ACM, 2016.
21. W. C. Yoon, S. Lee, and S. Lee. Burst analysis of text document for automatic concept map creation. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 407–416. Springer, 2014.
22. J. Zhang. The nature of external representations in problem solving. *Cognitive science*, 21(2):179–217, 1997.