Data-Driven Analyses of Electronic Text Books

Ahcène Boubekki[†], Ulf Kröhne[‡], Frank Goldhammer[‡],
Waltraud Schreiber[*], and Ulf Brefeld[†‡]

[†] Department of Computer Science, Technical University of Darmstadt, Germany
[‡] German Institute for Educational Research, Frankfurt am Main, Germany
[*] Faculty of History and Social Science, KU Eichstätt, Germany

Abstract. We present data-driven log file analyses of an electronic text book for history called the mBook to support teachers in preparing lessons for their students. We represent user sessions as contextualised Markov processes of user sessions and propose a probabilistic clustering using expectation maximisation to detect groups of similar (i) sessions and (ii) users. We compare our approach to a standard K-means clustering and report on findings that may have a direct impact on preparing and revising lessons.

1 Introduction

Electronic text books may offer a multitude of benefits to both teachers and students. They allow to combine text, images, interactive maps and audiovisual content in an appealing way, and the usage is supported by hyperlinks, search functions, and integrated glossaries. By representing learning content in various ways and enabling alternative trajectories of accessing learning objects, electronic text books offer great potentials for individualised teaching and learning. Although technological progress passed by schools for a long time, inexpensive electronic devices and handhelds have found their way into schools and are now deployed to complement traditional (paper-based) learning materials.

Particularly text books may benefit from cheap electronic devices. Electronic versions of text books may revolutionise rigour presentations of learning content by linking maps, animations, movies, and other multimedia content. However, these new degrees of freedom in presenting and combining learning materials may bring about also new challenges for teachers and learners. For instance, learners need to regulate and direct their learning process to a greater extent if there are many more options they can choose from. Thus, the ultimate goal is not only an enriched and more flexible presentation of the content but to effectively support teachers in preparing lessons and children in learning. To this end, not only the linkage encourages users to quickly jump through different chapters but intelligent components such as recommender systems [36] may highlight alternative pages of interest to the user. Unfortunately, little is known on the impact of these methods on learning as such and even little is known on how such electronic text books are used by students.

In this article, we present insights on the usage of an electronic text book for history called the mBook [37]. Among others, the book has been successfully deployed in the German-speaking Community of Belgium. We show how data-driven analyses may support history teachers in preparing their lessons and showcase possibilities for recommending resources to children. Our approach is twofold: Firstly, we analyse user sessions to find common behavioural patterns across children and their sessions. Secondly, we aggregate sessions belonging to the same user to identify similar types of users. This step could help to detect deviating learners requiring additional attention and instructional support.

In this paper, we argue that conclusions on an individual session or user basis can only be drawn by taking the respective population into account and propose a contextualised clustering of user sessions. We represent user sessions as fully observed Markov processes that are enriched by context variables such as timestamps and types of resources. We derive an expectation maximisation algorithm to group (user-aggregated) sessions according to the learners' behaviour when using the text book. To showcase the expressivity of our approach, we compare the results to a standard K-means-based [31] solution. While the latter leads to trivial and insignificant groups, our methodology allows to project similar sessions (users) onto arbitrary subsets of variables that can easily be visualised and interpreted. We report on observations that can be used to support teacher instructions and students learning.

The remainder is organised as follows. Section 2 reviews related work. We introduce the *mBook* in Section 3 and present our probabilistic model in Section 4. We report on empirical results in Section 5, Section 6 provides a discussion of the results and Section 7 concludes.

2 Related Work

The analysis of log files is common in computer science and widely used to understand user navigation on the web [25, 1]. Often, sequential approaches, such as Markov models and/or clustering techniques, are used to detect browsing patterns that are predictive for future events [35, 39, 18] or interests of the user [3]. However, previous approaches to modelling user interaction on the Web mainly focus on the pure sequence of page views or categories thereof, without taking contextual information into account. Patterns in page view sequences have been analysed using all sorts of techniques, including relational models [2], association rule mining [14, 15], higher-order Markov models [18], and k-nearest neighbours [6].

A useful step toward interpretable patterns is to partition navigation behaviour into several clusters, each with its own characteristics. Hobble and Zicari [24] use a hierarchical clustering to group website-visitors and Chevalier et al. [11] correlate navigation patterns with demographic information about users. Other heuristic approaches to identify clusterings of user interactions include sequence alignments [22], graph-mining [20]. The advantage of model-based clusterings is that the cluster parameters itself serve as a starting point for interpreting the results. Prior work in this direction focuses on modelling navigational sequences using Markov processes [10, 33] and hidden Markov models [40, 19, 8]. Haider et al. [21] cluster user sessions with Markov processes for Yahoo! News. Their approach is similar to ours as they also propose a nested EM algorithm, however, we model timestamps with periodic distributions while Haider et al. resort to simulate periodicity by external filtering processes. They focus on presenting clustering results and do not compare their methods to alternative ones.

In recent years, logfile analyses attract more and more researchers from other disciplines such as educational research [5]. Although the analysis of log and process related data is still a new and emerging field in educational research, two methodologies can be described [4]: *Educational data mining* (EDM) and *learning analytics*. Their common goal is to discover knowledge in educational data, however, the former is purely data-driven while the latter keeps the user/expert in the loop to guide the (semi-automatic) analysis. Many approaches are related to web-based learning environments such as MOOCs [9] and other learning management systems [32].

Köck and Paramythis [28] cluster learners in a web-based learning environment according to their performance in exercises and the usage of an interactive help. Lemieux et al. [30] develop an online exerciser for first year students and visualise identified patterns of usage and different behaviours. Sheard et al. [38] and Merceron and Yacef [34] analyse logfile data not only to describe user behaviour but also to provide information and feedback to learners and tutors. Generally, much work is put into visualising and organising the discovered knowledge as relationships and correlations are often complex and difficult to communicate [16].

Another line of research deals with the assessment of the level of motivation [23, 27]; the time spent on a page and filling in exercises turn out to be predictive indicators. Cocea [12] and Cocea and Weibelzahl [13] studied disengagement criteria as a counterpart of motivation. One of their findings showed that a user's history needs to be taken into account for predicting engagement/disengagement as exploratory phases always precede learning phases and vice versa.

3 The mBook

History as a subject is especially promising for a prototype of a multimedia textbook. The re-construction of past events and the de-construction of *historical narrations* is ubiquitously present in our historical consciousness; different narrations about the past contain and provoke a great deal of different intentions and interpretations. It is therefore crucial to deal with history in various and different perspectives as a single narration is always a retrospective, subjective, selective, and thus only a partial re-construction of the past.

The *mBook* is guided on a constructivist and instructional-driven design. Predominantly, the procedural model of historical thinking is implemented by a structural competence model that consists of four competence areas that are deduced from processes of historical thinking: (i) the competency of posing and answering historical questions, (ii) the competency of working with historical methodologies, and (iii) the competency of capturing history's potential for human orientation and identity. The fourth competency includes to acquire and apply historical terminologies, categories, and scripts and is best summarised as (iv) declarative, conceptual and procedural knowledge. It is often referred to as the foundation of structural historical thinking in terms of premises and results and systematised by principles, conceptions of identity and alterity as well as categories and methodological scripts [29].

Imparting knowledge in this understanding is therefore not about swotting historic facts but aims at fostering a reflected and (self-)reflexive way of dealing with our past. The underlying concept of the multimedia history schoolbook implements well-known postulations about self-directed learning process in practice. The use of the mBook allows an open-minded approach to history and fosters contextualised and detached views of our past (cf. [26]). To this end, it is crucial that a purely text-based narration is augmented with multimedia elements such as historic maps, pictures, audio and video tracks, etc. Additionally, the elements of the main narration are transparent to the learners. Learners quickly realise that the narration of the author of the mBook is also constructed, as the author reveals his or her construction principle.

The mBook consists of 5+1 chapters, *Antiquity*, *Middle Age*, *Renaissance*, *19th Century*, *20th and 21th Century* and a chapter on methods. In the German-speaking Community in Belgium, the mBook has about 1300 regular users. In our analysis, we focus on about 330.000 sessions collected in Belgium between March and November 2014 containing approximately 5 million events (clicks, scrolls, key press, etc.). The book encompasses 648 pages including 462 galleries and 531 exercises among others.

4 Methodology

4.1 Preliminaries

Let X denote the set of N user sessions given by $X = \{X^i\}_{i=1}^N$. A session is assembled by user events such as page views, clicks, scrolls, text edit, etc. In this work, we focus on the *connection time* t, the *sequence of the visited page* in terms of the chapter they belong to $\mathbf{x} = \langle x_1^i, \ldots, x_{T^i}^i \rangle$, and the *sequence of categories* realised by the viewed pages $\mathbf{c} = \langle c_1^i, \ldots, c_{T^i}^i \rangle$. The six chapters of the book together with the *homepage* and a *termination* page that encodes the end of a session form 8 possible realisations for every visited page, i.e., the values for the variable x_t^i. There are five different categories, *summary*, *text*, *gallery* and the auxiliary variables representing the categories for the *homepage* and the *termination* page.

4.2 Representation

We deploy a parameterised mixture model with K components to compute the probability of a session.

$$p(X^i|\Theta) = \sum_{k=1}^{K} \kappa_k p(X^i|\Theta_k).$$

The variables κ_k represent the probability that a *random* session is generated by the k-th component and also known as the *prior probability* for cluster k. The term $p(X^i|\Theta_k)$ is the *likelihood* of the session given that it belongs to cluster k with parameters Θ_k. Defining sessions in terms of time, chapters, and categories allows to assemble the likelihood of a session as

$$p(X^i|\Theta_k) = p(t^i|\beta_k)p(\mathbf{x}^i|\alpha_k)p(\mathbf{c}^i|\gamma_k).$$

The browsing process through chapters is modelled by a first-order Markov chain so that pages are addressed only by their chapter. We have,

$$p(\mathbf{x}|\alpha_k) = p(x_1|\alpha_k^{init}) \prod_{l=2}^{L} p(x_l|x_{l-1}, \alpha_k^{tr})$$

where $\alpha_k = (\alpha^{init}, \alpha^{tr})$ is split up into parameters α^{init} for the first page view and the transition parameters α^{tr} for the process.

The category model depends on the chapters as we aim to observe correlations between different types of pages. This may show for example whether galleries of some of the chapters are more often visited (and thus more attractive) than others and thus generate feedback for the teachers (e.g., to draw students attention to some neglected resources) and developers (e.g., to re-think the accessibility or even usefulness of resources). Categories are modelled by

$$p(\mathbf{c}|\gamma_k) = p(c_1|x_1, \gamma_k^{init}) \prod_{l=2}^{L} p(c_l|c_{l-1}, x_{l-1}, \gamma_k^{tr})$$

where again γ^{init} is used for the prior category and γ^{tr} for the subsequent transitions.

4.3 Modeling Time

The model for the connection times is inspired by the approach described in [21]. The goal is to project the continuous time space into a multinomial space to ease the estimation process. For this purpose, we introduce fixed unique and periodic components that serve as a new basis for generating continuous time events. To capture periodic behaviours, 90 time components are defined: 48 daily and 42 weekly components. The connection model is a multinomial law over each component with parameters $\beta_{k,j}^d$ and $\beta_{k,j}^w$ where j encodes the component and

d and w refer to a daily or weekly setting, respectively. The probability for a session to start at a certain time t is therefore given by

$$p(t|\beta_k) = \sum_{j=1}^{48} \beta_{k,j}^{d} d_j(t_d) + \sum_{j=1}^{42} \beta_{k,j}^{w} w_j(t_w).$$

The components are derived from the normal distribution. Periodic constraints are embedded in the probabilities, so that the density is composed with the tangent function which, besides of being periodic, conserves the symmetry. The generic form can be written as

$$p_{\mu,\sigma,T}(t) = \frac{1}{\mathrm{erfc}(\frac{1}{\sigma})T} \exp\left(-\frac{1+\tan^2(\frac{\pi}{T}(t-\mu))}{\sigma^2}\right).$$

The period is governed by the parameter T, and erfc is the complementary error function. The parameter μ represents the expectation similarly to the normal law. A component is said to recover a time slot if its density in this interval is higher than half of its maximum. This condition is parametrised by standard deviation σ. For a component covering time unit Δ we obtain $\sigma = \tan\left(\frac{\pi}{T}\Delta\right)/\sqrt{\log(2)}$.

The daily components are centred every 30 minutes, have a duration of 30 minutes, and the first component is centred at 12:00am. The weekly components are centred every four hours with a duration of four hours, and the first one is centred on Monday at 2:00am. This shift allows a synchronisation with the schools working hours, as we have a slice for the morning between 8:00am and 12:00pm, one for the afternoon between 12:00pm and 4:00pm and another one for the evening between 4:00pm and 8:00pm.

To capture the daily and weekly behaviors, connection times are considered modulo the 48 slices of a day (t_d) or the 336 slices of a week (t_w). The daily d_j and weekly w_j distributions are described as follows

$$\begin{aligned} d_j(t) =& \frac{1}{48\,\mathrm{erfc}(\frac{1}{\sigma_d})} \exp\left(-\frac{1+\tan^2\left(\frac{\pi}{48}(x-j)\right)}{\sigma_d^2}\right) \\ w_j(t) =& \frac{1}{336\,\mathrm{erfc}(\frac{1}{\sigma_w})} \exp\left(-\frac{1+\tan^2\left(\frac{\pi}{336}(x-4-8j)\right)}{\sigma_w^2}\right) \end{aligned}$$

using the variances $\sigma_d = \tan\left(\frac{\pi}{48}\frac{1}{2}\right)/\sqrt{\log(2)} \simeq 0.039$ and $\sigma_w = \tan\left(\frac{4\pi}{336}\right)/\sqrt{\log(2)} \simeq 0.045$.

4.4 Optimisation

Given our mixture model and assuming independence of the user sessions, the likelihood of the sessions is given by

$$p(X|\Theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \kappa_k p(t^i|\beta_k) p(\mathbf{x}^i|\alpha_k) p(\mathbf{c}^i|\mathbf{x}^i, \gamma_k).$$

The joint likelihood needs to be maximised with respect to the parameters $\Theta = (\pi_k, \beta_k, \alpha_k, \gamma_k)$. For computational reasons, we address the equivalent problem of maximising the log of the likelihood. The optimisation problem becomes

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \ \log p(X|\Theta).$$

We develop an expectation maximisation (EM)-like algorithm [17, 10]. Expectation maximisation is an iterative approach to approximate a local maximum from a given set of parameters. The procedure works in two steps: the expectation step (E-Step) computes the expectation of the objective function related to the problem from the actual set of parameters and deduces two temporary distributions of the Markov sequences over the cluster and the time components called *class-conditional probability distribution* and *time component-conditional probability distribution* denoted by $P_{i,k}$ and $Z^{\bullet}_{i,k,j}$, respectively,

$$\begin{aligned} P_{i,k} &= \frac{\kappa_k p(X^i|\Theta_k)}{\sum_{k'=1}^{K} \kappa_{k'} p(X^i|\Theta_{k'})} \\ Z^d_{i,k,j} &= \frac{\beta^d_{k,j} d_j(t^i_d)}{\sum_{j'=1}^{48} \beta^d_{k,j'} d_{j'}(t^i_d)} \\ Z^w_{i,k,j} &= \frac{\beta^w_{k,j} w_j(t^i_w)}{\sum_{j'=1}^{42} \beta^w_{k,j'} w_{j'}(t^i_w)} \end{aligned}$$

The maximisation step (M-Step) re-estimates the parameters from these proposal distributions to increase the value of the objective function and thus also the likelihood. We give the update formulas of four of the parameters, as they can be easily translated to the other parameters:

$$\begin{aligned} \kappa_k &= \frac{\sum_{i=1}^{N} P_{i,k}}{\sum_{k'=1}^{K} \sum_{i=1}^{N} P_{i,k'}} \\ \beta^d_{k,j} &= \frac{\sum_{i=1}^{N} Z^d_{i,k,j} P_{i,k}}{\sum_{j'=1}^{42} \sum_{i=1}^{N} Z^d_{i,k,j'} P_{i,k}} \\ \gamma^{init}_{k,g} &= \frac{\sum_{i=1}^{N} P_{i,k} \delta(x^i_1, g)}{\sum_{g'=1}^{5} \sum_{i=1}^{N} P_{i,k} \delta(x^i_1, g')} \\ \gamma^{tr}_{k,g,h} &= \frac{\sum_{i=1}^{N} P_{i,k} \eta_{g,h}(x^i)}{\sum_{h'=1}^{5} \sum_{i=1}^{N} P_{i,k} \eta_{g,h'}(x^i)}, \end{aligned}$$

where g and h take their values in the fives possible types of pages. The function $\delta(x^i_1, g)$ is the Kronecker delta that equals 1 if the two arguments are equal and 0 otherwise. The function $\eta_{g,h}(x^i)$ returns the number of transitions in session x^i from a page of type g to a page of type h.

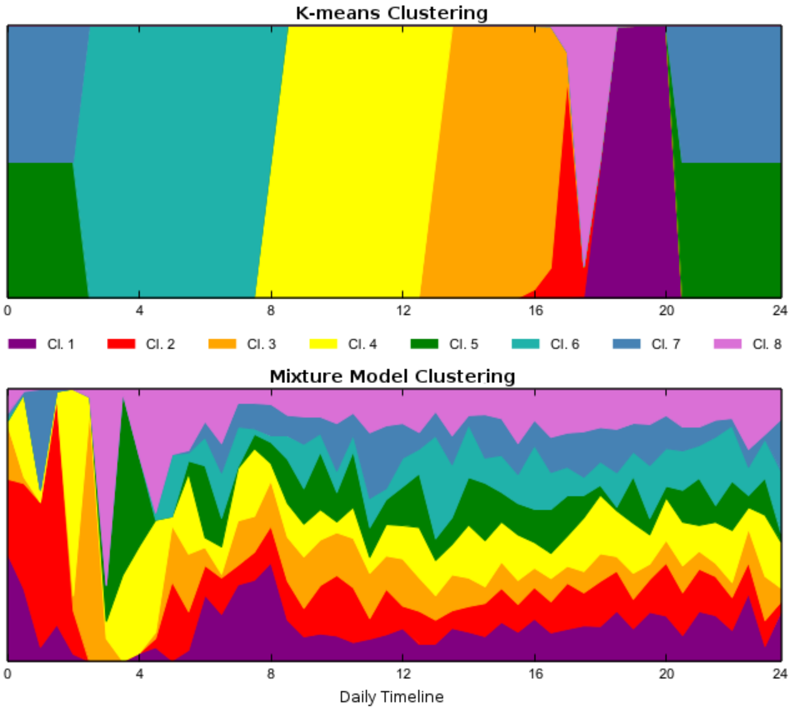

Fig. 1. Results for K-means and the proposed model.

5 Empirical Results

In our empirical analysis, we focus on about 330.000 sessions collected in Belgium between March and November 2014 containing approximately 5 million events including clicks, scrolls, key presses, etc. In the remainder, we show results for $K = 8$ clusters to trade-off expressivity and interpretability, however, other choices are possible.[1]

5.1 Comparison with K-Means

The first experiment demonstrates the expressivity of our approach. We compare our probabilistic solution with a winner-takes-all clustering by K-means [31]. Since K-means acts in vector spaces, user session are represented as vectors in a 354 dimensional space, so that both algorithms have access to identical information.

Figure 1 shows the results from clustering sessions. For lack of space, we focus on a projection of the final clusterings on the daily components capturing

[1] Note that we obtain similar results for all $1 \leq K \leq 30$; this holds in particular for the comparison with K-means.

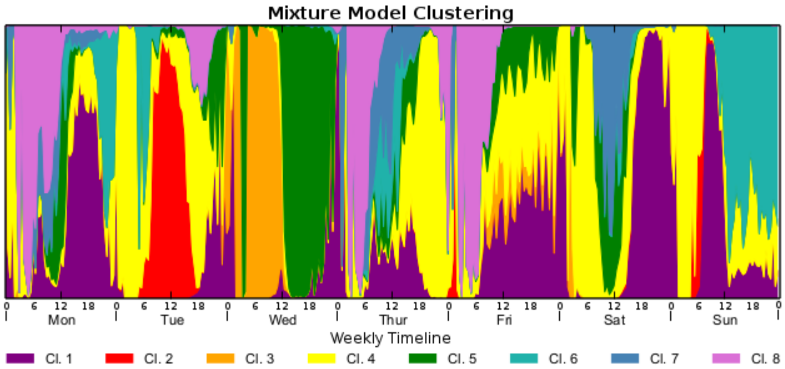

Fig. 2. Cluster distribution across a week.

repetitive behaviour across days. The geometric nature of K-means is clearly visible in Figure 1 (top): the clusters separate the day into six time slots of about four hours. The more complex colouring between 4pm and 8pm indicates that more variables than the connection time are active during that period. Nevertheless, the simplicity of the result (e.g., most of the clusters differ only in connection time), particularly for school hours, is clearly inappropriate for further processing or interpretation.

Figure 1 (bottom) shows the corresponding results for our probabilistic approach. The distribution of the clusters is fairly more interesting and balanced across the day. Clusters clearly specialise on dependencies across week days which is also shown in Figure 2. Cluster *C1* and *C4* capture recurrent behaviour and cover a large part of the user activity. Cluster *C6* focuses on the activity on Sunday afternoon and similarly, cluster *C5* specialises on Wednesday afternoon. Clusters *C2* and *C3* have a similar shape as they occur mainly on Tuesday and Wednesday morning, respectively, during school hours. In the remainder, we discard K-means and focus on the analysis of the proposed approach instead.

5.2 Session-based View

Figure 3 shows the results of a session-based clustering. User sessions are distributed across the whole clustering according to the expressed behaviour. Clusters can therefore be interpreted as similar user behaviours at similar times.

Before we go into details, recall Figure 2, where the Sunday afternoon is shared between cluster *C1*, *C4* and *C6*. The latter aggregates most of the activity and also most of the text page views. Figure 3 allows for a clearer view on the clustering. According to Figure 3, a similar observation can be made for clusters *C7* and *C8*. Both of them have *Antics* as the main chapter, and have a similar weekly distribution, the only difference being that *C8* contains more text views. The latter indicates more experienced users as we will discuss in the following.

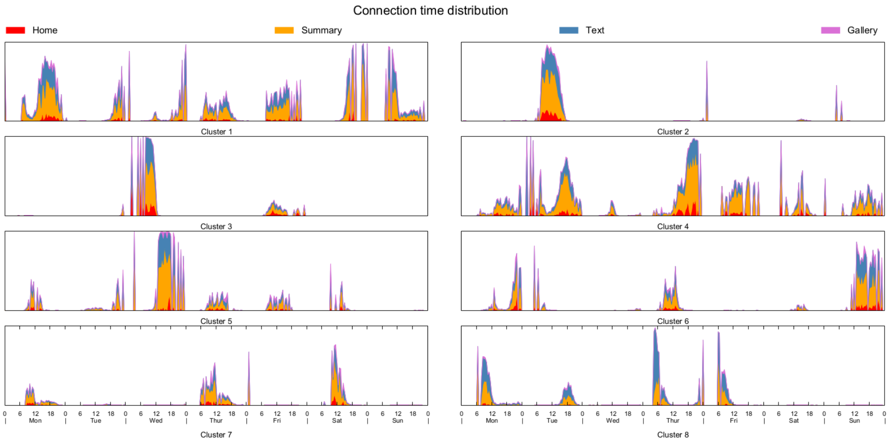

Fig. 3. Resulting clusters for the session-based clustering

The visualisation shows that all categories are clearly visible for all clusters, indicating a frequent usage of all possible types of resources by the users. Cluster *C6* possesses half of the mass on the weekend of category *text*. This indicates more experienced users who like to form their opinion themselves instead of going to summary pages. The same holds for cluster *C8* that possesses in addition only a vanishing proportion of the *home* category. Small probabilities of category *home* as well as large quantities of category *text* indicate that users continuously read pages and do not rely on the top-level menu for navigation.

5.3 User-based View

Our approach can also be used to group similar users. To this end, we change the expectation step of the algorithm so that sessions by the same user are processed together. That is, there is only a single expectation for the sessions being in one of the clusters. Clusters therefore encode similar users rather than similar behaviour as in the previous section.

Figure 4 shows the results. Apparently, the main difference of the clusters is the intensity of usage during working days and weekends. Cluster *C2* for instance clearly focuses on working day users who hardly work on weekends compared to Cluster *C1* whose users place a high emphasise on Saturdays and Sundays. Cluster *C3* contains low frequency users who rarely use the mBook and exhibit the smallest amount of sessions and page views per session (see also Figure 5). Cluster *C8* contains heavy (at night) users with high proportions of category *text*. In general, we note that transition matrices are consistent between chapters in contrast to the session-based clustering, that is, test takers interact with most of the chapters.

Figure 5 confirms our interpretations with descriptive statistics. Cluster *C8* containing the power users possess the highest number of sessions and also the

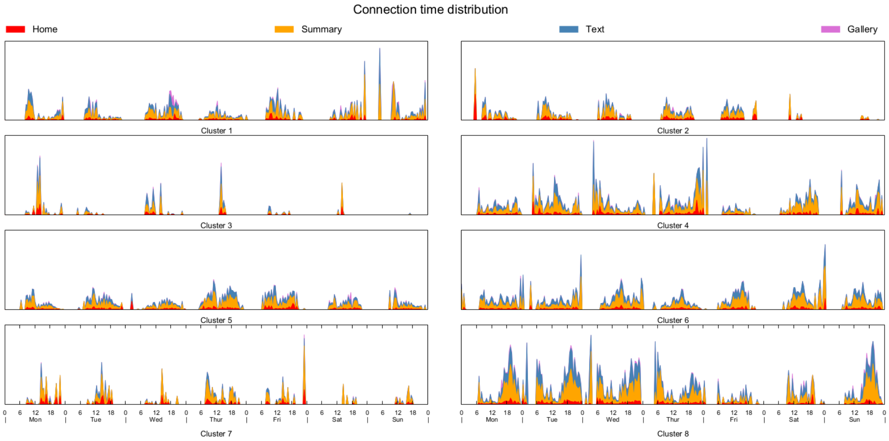

Fig. 4. Resulting clusters for the user-based clustering.

highest number of sessions per user. Clusters *C3* and *C7* are the smallest. As cluster *C3* has been identified as encoding low frequency users, it may be comforting to know that these users constitute a clear minority. Nevertheless teachers may be well advised to keep an eye on these children and individually support them by all means.

Figure 6 finally shows differences in clicking behaviour of users. For two clusters, transition matrices for types of resources are visualised, darker colours indicate more probable transitions. The two bars on top visualise the distribution of chapters. Both distributions are quite different. While users in cluster *C1* exhibit a broad interest and visit chapters uniformly, their peers in cluster *C4* focus clearly on the three chapters *Renaissance* and the *XIX* and *XX & XXI* centuries. However, the observation also shows that linkage is exploited by the users who seem to like browsing and learning about the book and thus also about history.

Users in clusters *C1* and *C4* exhibit very different click behaviours. While users in both clusters prefer viewing galleries, users in cluster *C1* move deterministically on to a text page, while users in cluster *C4* visit summary pages or terminate the session. This reflects the two possibilities of browsing inside the mBook: a hierarchical and a flat one. The former is realised by performing transitions between text pages always through a summary page. By contrast, a flat navigation makes use of the page to page navigation possibility and hence reflects the way a regular paper book is read. Knowing these relationships is an important means to personalise electronic books like the mBook. For instance, identifying an active session as a member of cluster *C1* allows to replace links to summary pages by other content as these users will almost always go back to a text page. On the other hand, knowing that a user of cluster *C4* is viewing

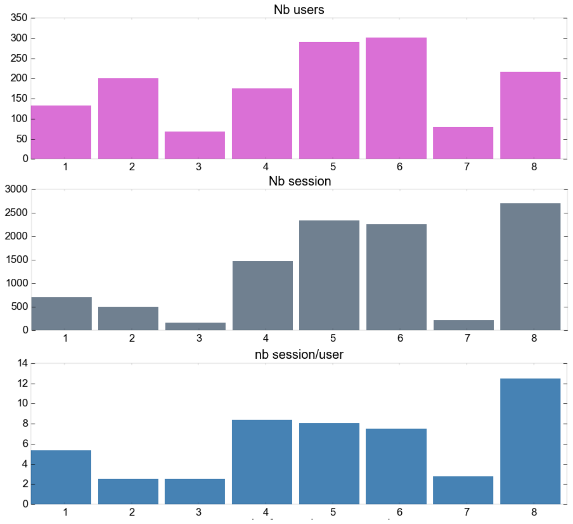

Fig. 5. Number of users (top), sessions (center), and sessions per user (bottom) for the user-view.

a gallery may be utilised to actively recommend other resources to prevent her from churning.

6 Discussion

Our results illustrate potential benefits from clustering learners for instructional purposes. In the first place, the probabilistic clustering approach shows a way how to condense a huge amount of logfile information to meaningful patterns of learner interaction. Classifying a student into one of several clusters reveals whether, when, and how the learner used the materials offered by the electronic text book. Thus, the teacher can get information about the learners' navigation speed, whether part of the content was used in self-directed learning processes as expected, whether learners came up with alternative learning trajectories, and so on and so forth. This information can be used by the teacher in a formative way (cf. the concept of formative assessment, e.g., [7]), that is, it is directly used to further shape the learning process of students. For instance, in a follow-up lesson the teacher could simply draw the students attention to some parts of the book that have not or only rarely been visited. Moreover, history and learning about history could be reflected in a group discussion of learners who used the mBook resources of a particular chapter in different ways.

An important extension of the presented analyses would be to relate contextual information (e.g., from teacher and class room level) to clusters. This would help to validate cluster solutions and improve their interpretability. For instance, a cluster of learners who use the text book mainly Thursday morning may con-

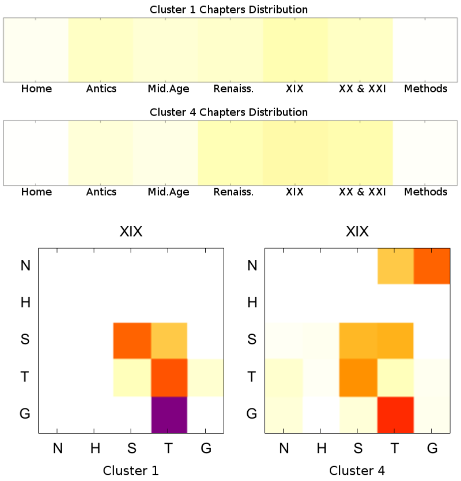

Fig. 6. Transitions (row → column) realised by two clusters.

sist of students with history lesson on Thursday morning and with teachers using the electronic text book only to support lessons and not for homework.

7 Conclusion

We presented contextualised Markov models to represent user sessions and proposed an Expectation Maximisation algorithm for optimisation. We applied our approach to clustering user sessions of the mBook, an electronic text book for history. Our results may have a direct impact on teachers and learners and can be used together with outlier analyses to find students who need individual support.

References

1. M. Agosti, F. Crivellari, and G. Di Nunzio. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, pages 1–34, 2011.
2. C. R. Anderson, P. Domingos, and D. S. Weld. Relational markov models and their application to adaptive web navigation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.
3. M. Armentano and A. Amandi. Modeling sequences of user actions for statistical goal recognition. *User Modeling and User-Adapted Interaction*, 22(3):281–311, 2012.
4. R. Baker and G. Siemens. *Educational data mining and learning analytics*. Cambridge Handbook of the Learning Sciences, 2014.
5. R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. 1(1):3–17, 2009.

6. D. Billsus and M.J. Pazzani. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2):147–180, 2000.
7. P. Black and D. Wiliam. Assessment and classroom learning. *Assessment in Education*, 5(1):7–74, 1998.
8. J. Borges and M. Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Transactions on Knowledge and Data Engineering*, pages 441–452, 2007.
9. C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong. Learning about social learning in moocs: From statistical analysis to generative model. Technical Report arXiv:1312.2159, 2013.
10. I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model-based clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.
11. K. Chevalier, C. Bothorel, and V. Corruble. Discovering rich navigation patterns on a web site. In *Proceedings of Discovery Science*, 2003.
12. M. Cocea. Can log files analysis estimate learner's level of motivation? In *Proceedings of the Workshop on Lernen - Wissensentdeckung - Adaptivität*, 2006.
13. M. Cocea and S. Weibelzahl. Log file analysis for disengagement detection in e-learning environments. *User Modeling and User-Adapted Interaction*, 19(4):341–385, 2009.
14. R. Daş and İ. Türkoğlu. Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications*, 36(3):6635–6644, 2009.
15. R. Daş and İ. Türkoğlu. Extraction of interesting patterns through association rule mining for improvement of website usability. *Istanbul University-Journal of Electrical & Electronics Engineering*, 9(18), 2010.
16. N. Delestre and N. Malandain. Analyse et représentation en deux dimensions de traces pour le suivi de l'apprenant. *Revue des Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF)*, 14, 2007.
17. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
18. M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. *ACM Transactions on Internet Technology (TOIT)*, 4(2):163–184, 2004.
19. M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. *ACM Transactions on Internet Technology*, 4(2):163–184, 2004.
20. Ş Gündüz and M. T. Özsu. A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2003.
21. P. Haider, L. Chiarandini, U. Brefeld, and A. Jaimes. Contextual models for user interaction on the web. In *ECML/PKDD Workshop on Mining and Exploiting Interpretable Local Patterns (I-PAT)*, 2012.
22. B. Hay, G. Wets, and K. Vanhoof. Mining navigation patterns using a sequence alignment method. *Knowledge and Information Systems*, 6:150–163, 2004.
23. A. Hershkovitz and R. Nachmias. Developing a log-based motivation measuring tool. In *Proceedings of the International Conference on Educational Data Mining*, 2008.

24. N. Hoebel and R. Zicari. On clustering visitors of a web site by behavior and interests. In *Advances in Intelligent Web Mastering*, volume 43, pages 160–167. Springer Berlin / Heidelberg, 2007.
25. B. J. Jansen. Understanding user-web interactions via web analytics. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–102, 2009.
26. Y. Karagiorgi and L. Symeou. Translating constructivism into instructional design: Potential and limitations. *Educational Technology and Society*, 8 (1):17–27, 2005.
27. J. Kay, N. Maisonneuve, K. Yacef, and O. Zaïane. Mining patterns of events in students' teamwork data. In *Proceedings of the ITS Workshop on Educational Data Mining*, 2006.
28. M. Köck and A. Paramythis. Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction*, 21(1-2):51–97, 2011.
29. A. Körber, W. Schreiber, and A. Schöner, editors. *Kompetenzen historischen Denkens: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik*. Neuried: Ars una, 2007.
30. F. Lemieux, M. C. Desmarais, and P. N. Robillard. Analyse chronologique des traces journalisées d'un guide d'étude pour apprentissage autonome. *Revue des Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF)*, 20, 2014.
31. S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
32. L. P. Macfadyen and S. Dawson. Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2):588–599, 2010.
33. E. Manavoglu, D. Pavlov, and C. L. Giles. Probabilistic user behavior models. In *Proceedings of the Third IEEE International Conference on Data Mining*, 2003.
34. Agathe Merceron and Kalina Yacef. A web-based tutoring tool with mining facilities to improve learning and teaching. In *11th International Conference on Artificial Intelligence in Education (AIED03)*, pages 201–208. IOS Press, 2003.
35. J. Qiqi, T. Chuan-Hoo, P. Chee Wei, and K. K. Wei. Using sequence analysis to classify web usage patterns across websites. In *Proceedings of the 45th Hawaii International Conference on System Science (HICSS)*, pages 3600–3609, 2012.
36. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.
37. W. Schreiber, F. Sochatzy, and M. Ventzke. Das multimediale schulbuch - kompetenzorientiert, individualisierbar und konstruktionstransparent. In W. Schreiber, A. Schöner, and F. Sochatzy, editors, *Analyse von Schulbüchern als Grundlage empirischer Geschichtsdidaktik*, pages 212–232. Kohlhammer, 2013.
38. J. Sheard, J. Ceddia, J. Hurst, and J. Tuovinen. Inferring student learning behaviour from website interactions: A usage analysis. *Education and Information Technologies*, 8(3):245–266, 2003.
39. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.
40. A. Ypma and T. Heskes. Automatic categorization of web pages and user clustering with mixtures of hidden markov models. In *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*, pages 35–49. Springer, 2003.