Browsing-Pattern Mining from e-Book Logs with Non-negative Matrix Factorization

Atsushi Shimada
Kyushu University
Fukuoka, Japan
atsushi@artsci.kyushu-u.ac.jp

Fumiya Okubo
Kyushu University
Fukuoka, Japan
fokubo@artsci.kyushu-u.ac.jp

Hiroaki Ogata
Kyushu University
Fukuoka, Japan
ogata@artsci.kyushu-u.ac.jp

ABSTRACT

In this paper, we report our work-in-progress study about browsing-pattern mining from e-Book logs based on non-negative matrix factorization (NMF). We applied NMF to an observation matrix with 21-page browsing logs of 110 students, and discovered five kinds of browsing patterns.

Keywords

e-Book logs, pattern mining, non-negative matrix factorization

1. INTRODUCTION

An e-Book system can collect various kinds of operation logs when a page is opened, when the next page is browsed and so on. The analysis of e-Book logs enables teachers to understand how a student browses a given material. However, just giving or showing the logs is insufficient to understand behaviors of students because of their diversity and high dimensionality. In this paper, we apply non-negative matrix factorization (NMF) technique [2], which is known as akin to principal component analysis and factor analysis. In [1], NMF is utilized to extract a Q-matrix[1] from observed test outcome data for n question items and m respondents. In our study, we discover students' browsing patterns, i.e., how they browsed the given material, from e-book logs data for n page browse and m students. Besides, we analyze the relationship between the patterns and quiz scores.

2. E-BOOK LOGS

The e-Book logs were collected from 110 first-year students in an information science course taken in the first semester of the 2015 school year at Kyushu University in Fukuoka, Japan, via BookLooper (Kyocera Maruzen Systems Integration Co., Ltd.). Figure 1 shows samples of e-Book logs. There are many types of operations in logs, for example, OPEN means that the student opened the e-book file and

[1]A mapping of item to skills is termed a Q-matrix

User	Material	Operation	PageNo	Date	Time
X	00000000NLAT	OPEN	0	2014/10/15	9:01:09
X	00000000NLAT	CLOSE	1	2014/10/15	9:01:13
Y	00000000P82P	PREV	25	2014/10/29	10:05:35
Y	00000000P855	NEXT	2	2014/11/19	8:52:47
Z	00000000P84Z	NEXT	9	2014/11/12	9:31:30
...	...	...	...	...	...

Figure 1: Samples of e-Book logs

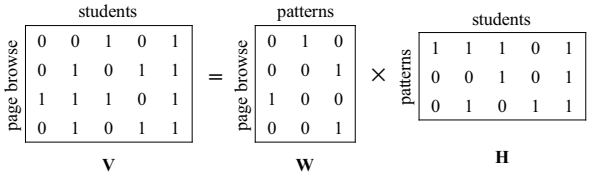

Figure 2: Pattern mining from browsing matrix

NEXT means that the student clicked the next button to move to the subsequent page. The duration of browsing each page can be calculated by subtracting the timestamps between subsequent pages.

3. METHODS

We utilize non-negative matrix factorization (NMF) technique to discover some browsing patterns. NMF approximately decomposes a matrix of $n \times m$ positive numbers V as the product of two matrices:

$$V \approx WH. \tag{1}$$

NMF imposes the constraint that the two matrices, W and H, be non-negative. In our approach, the matrix V, named browsing matrix, is represented by the fact whether a student browsed a page or not. More specifically, we set an element $v_{i,j}$ of the matrix V by

$$v_{i,j} = \begin{cases} 1 & (\text{if } t_{i,j} > th) \\ 0 & (\text{otherwise}), \end{cases} \tag{2}$$

where $t_{i,j}$ is the duration of page i browsed by student j. The decomposed matrices represent two latent relationships: "page browse vs. patterns" given by matrix W and "patterns vs, students" given by matrix H. In the sample of Figure 2,

Table 1: Description of discovered browsing patterns

pattern 1	browse the latter part of pages
pattern 2	browse the former part of pages
pattern 3	browse the middle part of pages
pattern 4	browse the beginning and end part of pages
pattern 5	browse pages between #12 and #15

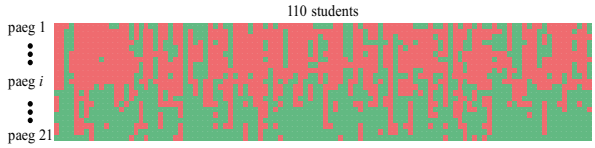

Figure 3: Browsing matrix. The red and green colors denote the value of $v_{i,j}$, red for one, green for zero, where the th was set to be 10 seconds.

browsing patterns are represented by three patterns in W. Meanwhile, H means whether a student has one or more browsing-patterns for a given material. In the experiments, we set the number of patterns to be five.

4. EXPERIMENTS

The browsing matrix used in our experiments were obtained from 110 first-year students. The students were asked to preview the material in advance before the lecture. They browsed the given material of information science which consists of 21 pages with a spread display setting. Therefore, the V is represented by 21-row $\times$ 110-column matrix as shown in Figure 3. The column of V corresponds to a student's previewing history whether he/she spent time at page i longer than th or not, which is calculated by formula (2). In the experiment, we set the th to be 10 (second).

NMF was performed to find five patterns. The decomposed matrices W and H are shown in Figure 4 and Figure 5 respectively. Note that the W is transposed in the figure due to the limitation of page space. Each pattern can be roughly described as Table 1.

The upper part of Figure 5 shows the correspondence between a student and his/her browsing pattern. The red color means that the student has the pattern (for example, the student in the most left column has pattern 2 and pattern 4). After the NMF, we acquired five groups based on consensus clustering technique (refer to literature [3] for more details). The bottom part of Figure 5 is the reordered matrix of W to show the group characteristics efficiently. The group 1 has the pattern 2 more strongly than the other patterns, which means that they spent longer time on the former part of pages.

We compared the student groups with their quiz scores (see Table 2). The quiz (max score = 10) was conducted at the beginning of the lecture. The average score of group 4 was lower than the other groups because the group had no pattern, i.e., they did not browse the material well. On the other hand, the group 2 got the highest score. They had the pattern 5, which corresponds to browsing the pages between #12 and #15. The contents of these pages were related

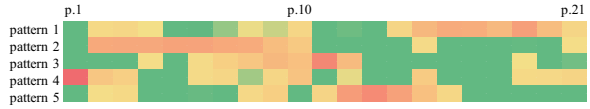

Figure 4: Visualized matrix W. Red parts represent the correspondence to each pattern. For example, pattern 2 denotes that pages from 2 to 10 are well browsed.

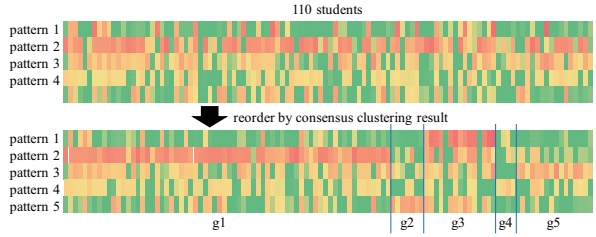

Figure 5: (Top:) Skill matrix visualized by color scale. The red color represents larger value. (Bottom:) Reordered pattern matrix based on consensus clustering result. There are five groups (g1, ..., g5) found by clustering.

Table 2: Average scores of quiz in each student group

student group	g1	g2	g3	g4	g5
average score	6.25	6.95	6.57	5.49	6.00

to the practice exercise to enrich the understanding. We guess that the students in group 2 could work the exercise because they had already understood the basic contents in the material. Therefore they got better quiz scores than the other student groups.

5. CONCLUSION

In this paper, we gave our work-in-progress report about e-Book browsing pattern mining and its potentials to fathom the relationships between patterns and understanding level of contents. In the experiments, we showed a primal result of pattern mining based on NMF. We found out that NMF could provide reasonable decomposed matrices to explain the browsing patterns. In the future work, we investigate the appropriate number of patterns because we predefined the number of patterns in this paper. Besides, we have to consider more effective method to generate a browsing matrix from e-Book logs.

6. REFERENCES

[1] M. Desmarais. Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In Proceedings of the 4th International Conference on Educational Data Mining, pages 41–50, 2011.

[2] D. Lee and H. Seung. Learning of the parts of objects by non-negative matrix factorization. Nature, 401:788–791, 1999.

[3] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn., 52(1-2):91–118, July 2003.