Monitoring User Pattern in School Information System Using Logfile Analysis

Arne Hendrik Schulz and Andreas Breiter

Institute for Information Management, University of Bremen, Germany
ahschulz@ifib.de

Abstract. Analyzing user patterns in school information systems can be difficult as several methods (e.g. interviews, surveys, and observations) can be time-consuming. We propose logfile analysis as a method that offers several advantages, primarily non-reactive data capture. With logfiles from a school with over 100 teachers over a seven month period, we try to get a deeper insight about the system's usage and the interactions between users. The results show that three user groups can be identified, classified by the intensity of usage. Network graphs helped us to visualize a complex system and helped us to identify important subjects and categories. Nevertheless, logfiles alone lack in providing information giving deeper insights about uses of the system like user goals and aims.

Keywords. Data mining; web mining; school information systems; logfile analysis.

1 Introduction

School information systems (SIS, see [1]) are increasingly used in schools for teachers to communicate with each other, with pupils and with parents. The systems' functions include resource planning, groupware, budget and decision support. Most research in this field is concentrated on interactions between teachers and pupils. Empirical studies on uses of learning management systems (especially in higher education) have shown interesting results about the teaching and learning process [2, 3].

In contrast, our research focus is the communication between teachers and school management (e.g. the principal). The research project "Mediatized organization worlds in schools: Schools as translocal network organizations", funded by the German Research Foundation, focuses on the communication between school members within school information systems [4]. The central research question goes beyond the assumption of one media logic [5] as it tackles the interdependence between media change and organizational change in one specific social world. The empirical research is based on a triangulation of three methods: participant observation within schools; group discussion with teachers and senior staff; and logfile analysis of school information systems.

This paper will focus on the logfiles of one SIS, used to uncover teachers' behavior within these systems and to discuss the benefits and limits of the method critically.

2 Logfile analysis

Logfiles have their origin in the technical basis of server-based software systems – they provide information, problems or errors about the systems and its applications [6, 7]. Programmers use logfiles also for debugging. The term “logfile” is often associated with webserver logfiles like the Extended Common Logfile Format [8]:

1.2.3.4 - - [25/Aug/2011:12:15:33 +0100] "GET /index.php HTTP/1.1" 200 23578 - "Webbrowser (System etc.)"
1.2.3.4 - - [25/Aug/2011:12:15:47 +0100] "GET /page2.php HTTP/1.1" 200 15789 "http://www.domain.com/index.php" "Webbrowser (System etc.)"

Looking at these fictitious logfiles from a webserver (see above), we can identify the user by her Internet Protocol (IP) (1.2.3.4) and additionally the Browser-Operating System (OS)-combination if multiple users use the same internet connection. If a user is identified, one can track the movement within the site because a second last entry contains the page the user came from, the so-called referer. In the example above, the user enters the site at ‘index.php’, stays on the site for 14 seconds and moves on to ‘page2.php’ by using a hyperlink. These “clicks” are called actions. Using this information, we can track movements from all users separately.

Nowadays, marketing [9] and (e-)commerce [10] also use the advantages of analyzing logfiles using data mining methods [9, 11]. Research tries, for example, to identify usage patterns of a website and to take the results as a basis for optimizing the website’s structure and to show advertisements. Another aim is to work with automated recommendation systems [12, 13].

Working with logfiles is summed up under the term web usage mining [14 - 16], a subcategory of data mining [17, 18], which aims to transfer existing data mining methods to the field of the internet or web. Practically, there are mainly five ways used by other researchers to conduct logfile analysis:

1. A descriptive analysis to display which pages of a website are accessed more than others and how many users selected a specific function, e.g. the search function [19]. Many free and commercial logfile tools have these capabilities too.
2. Besides descriptive analysis, logfiles are used to show paths from users or visitors through the site. One often-used algorithm here is sequential patterns [20], for example [12, 21 - 24].
3. Logfiles are also used to cluster users or visitors into groups. The clusters are based on movements or paths through the system.
4. Social network analysis [25 - 27] to display connections between users and/or websites are based on the “clickstream”-data [18].
5. Adapting other statistical methods and algorithms (e.g. multilevel analysis) for logfile analysis.

In contrast to other methods, logfile analyses have the advantage of being non-reactive. All information is gathered on the application layer or server layer and not registered by the user. Furthermore, the data are stored in a machine-readable format and can be used immediately. The main disadvantage, from the researcher’s point of

view, is that one has no information about the user's intentions and goals. Furthermore, there is usually no information about the user itself like gender, age, etc. The logfiles themselves only show the users' behavior within the system. This applies especially to the last action within the system as we do not know if the user reached her goal or not.

Beside these problems, uses of logfiles can lead to high privacy concerns. The users normally have no control over the logfiles that are produced by the server or application. Therefore, logfiles must be made anonymous by researchers.

3 The School Information System Used

The research project was conducted in two schools in German cities with more than 100,000 inhabitants. Each school had more than 100 teachers and more than 1,000 pupils. The results show the behavior in one of the two schools. The School Information System (SIS) was used among the staff to coordinate and communicate with each other. It was hosted by an external company and had limited administrative effort need in the school. Additionally, it had the advantage that the SIS could be accessed not only from inside the school, but also from home.

The SIS offered the following opportunities for the staff to communicate among each other:

- Announcements.
- Dates.
- Materials and files.
- Discussions.

The system also offered a Learn Management System (LMS) for teachers and classes. Users entered the system via a fixed Uniform Resource Locator (URL). Changes could also be followed via Really Simple Syndication (RSS)-feeds, dates could be subscribed via iCal and therefore be used with Smartphones and extended groupware-programs (Outlook, iCal for Mac, Mozilla Thunderbird).

The system's structure was somewhat different from normally-used information systems. These tend to use a hierarchical structure like classes within grades or classes within subjects. This system offered a much flatter system with intense usage of categories and keywords. Every item (announcement, date, etc.) could be put in several categories and could be tagged. This was especially important for working with materials.

Users needed to work with categories and keywords or use the search function to find the relevant materials. All materials could only be sorted by name, creation date and user. This meant that materials had to be tagged to ensure other users could find them as scrolling and searching manually was ineffective. So, one material could have more than one file.

Announcements did not rely on correct categorizing and tagging as there were not many new entries. The overview page (which was similar to the overview page of the materials) showed the main information and usually all new entries. Dates were not dependent on correct tagging and categorizing either, as the system offered a calendar

view that showed all dates, ordered by month, week or day. As mentioned above, dates could also be accessed with handhelds or Smartphones that also did not rely on keywords and categories.

The analyzed logfiles ran from March 2011 to March 2012, with some interruptions, especially with no entries from the end of April 2011 to the middle of June 2011. We had a total of 120,000 hits during the whole period from 138 users. After the deletion of all iCal and RSS accesses and path completions, the sum of hits was about 62,000. The 138 unique users had a total of 4,451 visits (a visit defined as a sequence of hits from a unique user, ending after 30 minutes of inactivity [28]).

4 Preliminary Outcomes

We analyzed the logfiles in different ways and took the five afore-mentioned ways in section 2 as a basis. At first, we looked at the descriptive statistics of the users and their access patterns in the SIS and identified three groups: heavy; medium; and minor users. There is a smooth transition between minor and medium users and a bigger gap between medium and heavy users (see Figure 1). Minor users have between 0 and 17 visits in the monitored period, and medium users between 18 and 200 visits. Most users of the SIS (a total of 138) were classified in the minor group (79), with about 40% in the medium group (56) and only three users in the heavy users group. We assumed that the three heavy users belonged to senior staff. Observations and group discussions supported this assumption, but as logfiles were taken anonymously, we could not be sure.

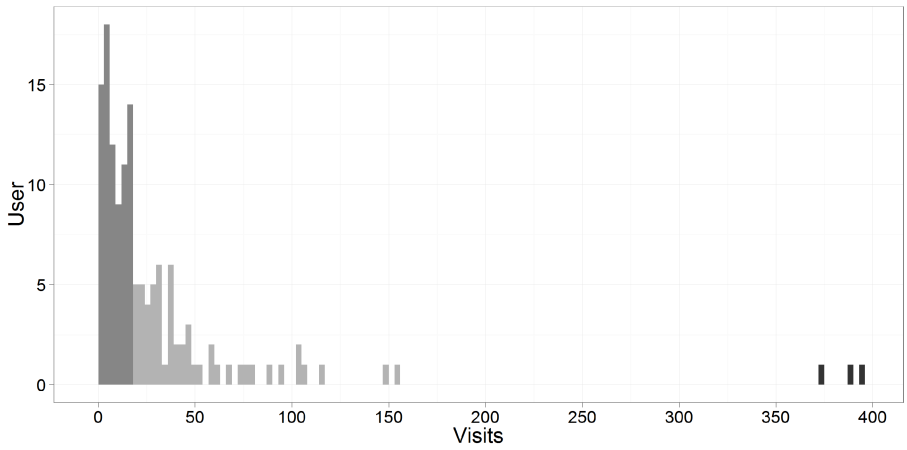

Figure 1. Different groups according to visits.

The SIS offers a variety of collaboration options, e.g. announcements, invitations, materials and discussions. Materials and dates are accessed most frequently, followed by announcements. Teachers accessed the SIS especially from Sunday until Tuesday, and slightly less on Wednesdays and Thursdays (see Figure 2).

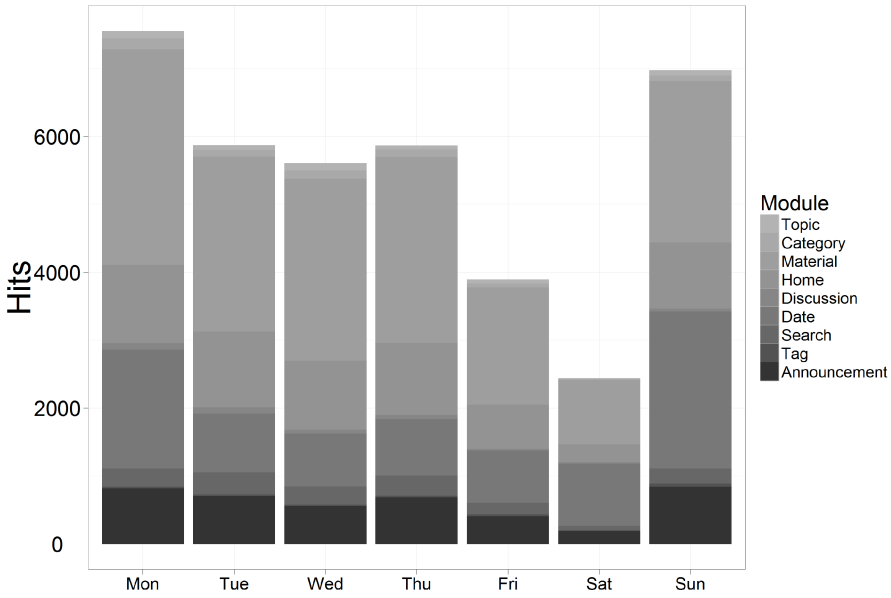

Figure 2. SIS access by weekday and modules.

After looking at descriptive statistics, we took a deeper look at user patterns within the system. Using web mining techniques [17] we transformed the logfiles into real clickstream data, which enabled us to follow the navigation from users within the system. In detail, we worked with the accessed URL in the logfile and the accessed URL in the next line (see the logfile examples above). We used these data to draw social networks with items as nodes and actions from one item to another as edges. This step offered us several advantages. Firstly, we got an overview of the complex system. All accessed items were displayed on the network map arranged using force-based algorithms [29], which meant that nodes connected through frequently-used navigational paths were displayed closer together (and pages containing user information were excluded from further analysis). This also implied that we did not recognize any items that were not visited. Secondly, the edges indicated "movement" between items and showed strong relations between two or more nodes and highlighted central nodes (gatekeepers). One outcome for example was that tags, groups and school subjects (beneath index-pages) were mainly used to navigate through the SIS. Within subjects, linked items were also one main source of navigation. Social network analysis software offered the possibility to color the nodes by their type (material, date, announcement, tag, etc.) and model their size by using visit numbers. This meant that thicker nodes had more visits than thinner ones. Using these features, we could display this outcome in a very comfortable way.

The network graphics in Figure 3 show the analyzed SIS as a whole. There are three eye-catching groups in the upper part of the graph, each one connected to one category. These categories are (from left to right): miscellaneous; reports; and confer-

ences. All are mainly linked to dates, some announcements and materials. Announcements and materials are more likely to be accessed than dates. This is no surprise as dates can be viewed in a calendar-like overview. The items themselves are mainly linked to the category and not linked among themselves.

In the bottom left are many materials closely connected to each other. Above these materials are the two subjects English (bigger) and Spanish (smaller). In contrast to the representation of the former three categories, the nodes are overlapping each other and are not only linked to the subject itself but also to each other. This indicates that the items are closely linked together. The relative big node size is another indicator for the higher level of material exchange within these two subjects.

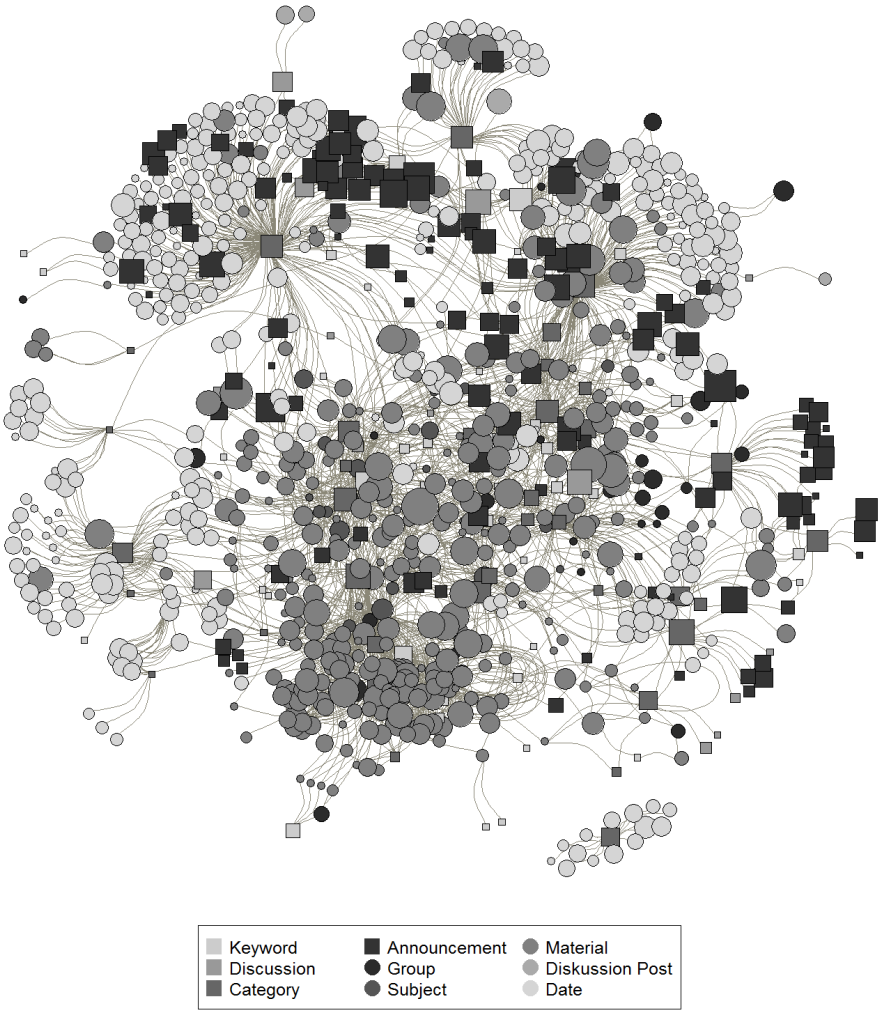

Figure 3. SIS represented as a network graph.

In the middle of the graph more subjects are to be found, but the representation does not highlight them. Above English and Spanish is a stronger concentration of subjects, which include mainly STEM subjects (science, technology, engineering and mathematics). All subjects have connections to materials, categories and keywords, but as said before, no subject is eye-catching. But they are connected to some of the most accessed materials in the system, which makes it important to take a closer look at all subjects, not only English and Spanish.

Figure 4 shows a scatter plot with number of materials per subject and the sum of hits of these materials. The plot shows only materials created within the logged period. Previously-created materials and materials that do not have a direct link to teaching were eliminated. The size of each subject shows the number of different contributors. The plot shows that English has the most hits (2,300) and the most new materials (23). Seven teachers contributed materials to the subject. That is no surprise and was already assumed. Spanish, on the other hand, is more interesting. It has the second most new materials (15), but only around 500 hits and only three contributors. Social science, the subject with the third most materials (8) has 1,500 hits on those materials. It has also more contributors (6). Other subjects like mathematics, German, business studies and chemistry have all less than five new materials during the period, but more accesses than Spanish. One explanation may be that there are more teachers in English and STEM subjects than in Spanish. Having only three teachers that are exchanging files throughout the system in a language that not many other teachers in Germany understand and speak (in contrast to English), means that it may depend on a small focus group of these materials. Teachers from other subjects may also not be interested in these materials due to the language barrier.

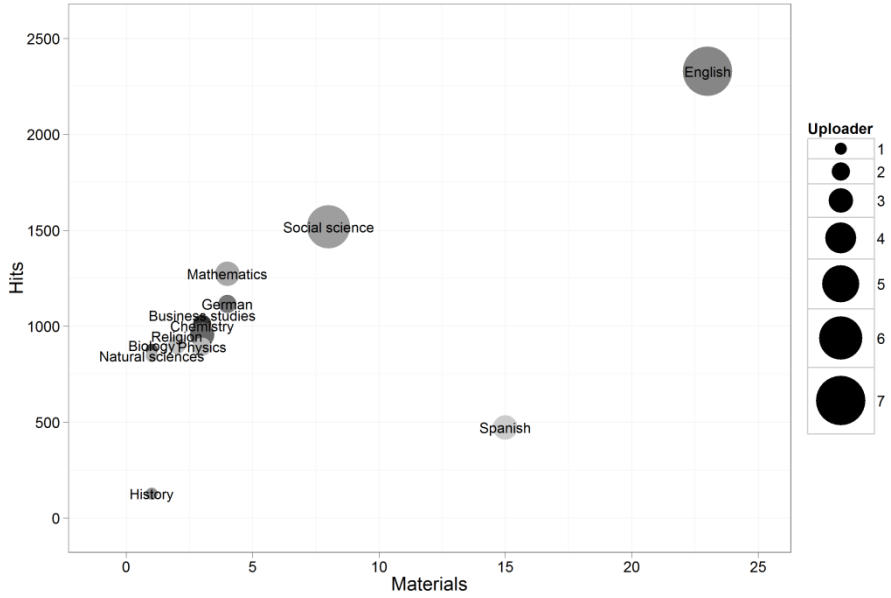

Figure 4. Materials and accesses by subject.

5 Limitations of logfile analysis

In general, logfiles can give researchers a useful view into a system and its usage, as shown in the previous section. Their technical background offers some neat advantages. Firstly, logfiles are already stored in a machine-readable file with a predefined format (e.g. Extended Common Logfile Format). This offers easy data management and also easy appending of new logfiles. Secondly, logfiles are usually enabled in web-based systems and can easily be adjusted to researchers' needs. The greatest advantage of this method is their non-reactivity. The knowledge of being supervised may lead to unusual behavior, but in logfile analysis, the users do not know that their actions are monitored.

This leads of course to ethical and privacy concerns. Ethical concerns may be solved by informing the users afterwards that they have been monitored and remove those users who disagree to the method. Privacy concerns are often connected to privacy laws, which, for example, prohibit the identification of users. Providers truncate IP addresses or identifier fields sometimes in order to fulfill these requirements, which may hinder logfile analysis.

But there are three more main disadvantages we discovered in logfile analysis:

1. *Combination with other data:* Due to privacy laws, users can usually only be identified by the IP address (and Browser-OS field) or an extra identifier field. These identifiers cannot be combined with additional data like gender, age, position, income or the subject in which a teacher gives lessons. Even if users accept the combination of different data sources, it may be impossible to do so as they usually do not know their internal user identifier or IP address.
2. *Identifying significant behavior:* The analyzed data had a time span of about 300 days. There may be the possibility to overlook significant behavior as the amount of data is large and significant behavior might not be the most common behavior. Additionally, commonly used statistical methods like sequential patterning or cluster analysis try to find common and frequent patterns, not rare or unique patterns. This may lead to a mismatch between available methods and research aims. However, the (normal) size of logfiles requires quantitative methods and cannot be analyzed by manual methods like inspections.
3. *User aims:* User paths or actions do not automatically reveal user aims. This applies especially to the last action of a visit. There is generally no information about the aim (the reason for using the system) itself and if the aim was reached by the user. All accessed items of a system (e.g. material, date or announcement) during a visit may also not indicate what a user was searching for or planning to access. The user could just drift through the system or check all new items.

These problems restrict the (scientific) outcome or results of logfile analysis, especially when working with assumptions and hypotheses. As long as the data cannot be combined with more information that reveals the user's aims for using the system, it is not easy to explain more about the usage.

6 Conclusions

Logfile analyses allow the researcher a deep insight into user patterns within a SIS. Furthermore, the data collection is easy and non-reactive. Initial results show a diversity of the system's usage with respect to users, items and actions. To identify these, we conducted descriptive analysis and social network analysis.

Currently we are working on a next step, clustering users to profiles. This will help to identify typical usage patterns within the SIS. As logfiles are anonymous, this can help us identify main actors within the system and help us to eventually diversify the above-mentioned three groups. Web usage mining might also help us to better identify important items within the system. We conducted multilevel analysis [30] to find out which characteristics of an item led to higher requests. Results indicate that the access to a new item depends on the item itself, its type and (of course) the time elapsed since its creation. The user who created the item also plays an important role.

Nevertheless, anonymous logfiles limit us in revealing more about the user aims. We therefore do not only use logfile analysis to test the hypotheses of our project, but also rely on participant observations and group discussions. This combination, for example, helped us to gain more information about the teachers' behavior concerning the uploading and accessing of materials of different subjects. During the participant observation, we discovered that English teachers are using the system quite often and group discussions revealed that one Spanish teacher started uploading materials into the SIS and two other teachers followed this behavior during the research period.

References

1. Breiter, A., Lange, A., Stauke, E. (eds.): School Information Systems and Data-based Decision-Making. Schulinformationssysteme und datengestützte Entscheidungsprozesse. Peter Lang, Frankfurt-am-Main, Germany (2008)
2. Hew, K.F., Brush, T.: Integrating Technology into K-12 Teaching and Learning: Current Knowledge Gaps and Recommendations for Future Research. Educational Technology Research and Development, 55 (3), 223–252, (2007)
3. Pelgrum, W.J.: Obstacles to the integration of ICT in education: results from a worldwide educational assessment. Computers & Education, 37 (2), 163–178, (2001)
4. Breiter, A., Welling, S., Schulz, A.H.: Mediatisierung schulischer Organisationskulturen. In Hepp, A., Krotz, F. (eds.): Mediatisierte Welten: Beschreibungsansätze und Forschungsfelder. VS Verlag, Wiesbaden, pp. 96–117, (2011)
5. Schulz, W.: Reconstructing Mediatization as an Analytical Concept. European Journal of Communication, 19 (1), 87–101, (2004)
6. Oliner, A., Ganapathi, A., Xu, W.: Advances and challenges in log analysis. Communications of the ACM, 55 (2), 55–61, (2012)
7. Suneetha, K.R., Krishnamoorthi, R.: Identifying User Behavior by Analyzing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security, 9 (4), 327–332, (2009)
8. Markov, Z., Larose, D.T.: Data mining the web: uncovering patterns in web content, structure and usage. Wiley, Hoboken, NJ (2007)

9. Büchner, A.G., Mulvenna, M.D.: Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. ACM SIGMOD Record, 27 (4), 54–61, (1998)
10. Song, Q., Shepperd, M.: Mining web browsing patterns for E-commerce. Computers in Industry, 57 (7), 622–630, (2006)
11. Larose, D.T.: Data Mining Methods and Models. Wiley-IEEE Press, Hoboken, NJ (2006)
12. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. Communications of the ACM, 43, 142–151, (2000)
13. Zaiane, O.R., Srivastava, J., Masand, B., Spiliopoulou, M. (eds.): WEBKDD 2002 - Mining web data for discovering usage patterns and profiles. 4th international workshop, Edmonton, Canada, July 23, 2002. Revised papers, Springer, Berlin, Germany (2003)
14. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. IEEE International Conference on Tools with Artificial Intelligence, IEEE Computer Society 558, (1997)
15. Huang, X., An, A., Liu, Y.: Web Usage Mining with Web Logs. In Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining, Second Edition. Information Science Reference, Hershey, PA, pp. 2096–2102, (2008)
16. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor Newsl, 1 (2), 12–23, (2000)
17. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Springer, Berlin, Germany (2011)
18. Mobasher, B.: Web Mining Overview. In Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining, Second Edition. Information Science Reference, Hershey, PA, pp. 2085–2089, (2008)
19. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. Wiley, Chichester (2010)
20. Agrawal, R., Srikant, R.: Mining Sequential Patterns. Proceedings of the Eleventh International Conference on Data Engineering, 3–14, (1995)
21. Banerjee, A., Ghosh, J.: Clickstream Clustering using Weighted Longest Common Subsequences. Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining (2001)
22. Chen, J., Cook, T.: Mining contiguous sequential patterns from web logs. Proceedings of the 16th international conference on World Wide Web, ACM, pp. 1177–1178, (2007)
23. Hay, B., Wets, G., Vanhoof, K.: Mining navigation patterns using a sequence alignment method. Knowledge and Information Systems, 6 (2), 150–163, (2004)
24. Mannila, H., Toivonen, H., Verkamo, A.I.: Discovering frequent episodes in sequences (Extended Abstract). 1st Conference on Knowledge Discovery and Data Mining, pp. 210–215, (1995)
25. van der Aalst, W.M.P., Reijers, H.A., Song, M.: Discovering Social Networks from Event Logs. Computer Supported Cooperative Work, 14 (6), 549-593, (2005)
26. Hogan, B.: Analyzing Social Networks via the Internet. In Fielding, N., Lee, R.M., Blank, G. (eds.) The SAGE handbook of online research methods. SAGE, Los Angeles, CA, pp. 141–160, (2008)
27. Wasserman, S., Faust, K.: Social network analysis : methods and applications. Cambridge University Press, Cambridge, NY (1994)
28. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide web. Computer Networks and ISDN Systems, 27 (6), 1065–1073, (1995)
29. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and experience, 21 (11), 1129–1164, (1991)

30. Twisk, J.W.R.: Applied Multilevel Analysis. Cambridge University Press, Cambridge (2007)