Masaryk University
Faculty of Informatics

Web Usage Mining on is.muni.cz

Master's Thesis

Lukáš Čenovský

2003

Assignment

- Web usage mining overview.
- Application of web usage mining algorithms and procedures on the log data from is.muni.cz server. The goal is to analyze the user's browsing paths, find unexpected rules and patterns, discover common sequential patterns for groups of users, detect differences in sequential patterns during different time periods (e. g. enrollment, exam session, holidays).
- Advice to improve is.muni.cz service.

Declaration

I hereby declare that this thesis is my original work. I cite all the sources used in this work together with reference to the corresponding source.

Acknowledgment

I would like to thank Ing. Michal Brandejs, Václav Lorenc, Petr Šmejkal, and PhDr. Ivana Tulajová for the contributed help.

Abstract

Web Usage Mining is the foundation for a Web site analysis. It employs various knowledge discovery methods to gain Web usage patterns. This thesis describes the Web Usage Mining procedures and methods in the first part (Chapter Three). The second part is focused on our practical tests on the data from the Information System (IS) of Masaryk University (Chapter Five). We focused on the statistical overview of the IS usage during one semester. Then, we chose four scripts on which we performed the path analysis. We present the usual departure and the following page of these scripts, and the most popular trails between them and one more script.

Keywords: data mining, web usage mining, information system

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information dominates the world more than anytime before. We live in the information world. The Internet is the roads and the highways in this world, the content providers are the road workers, and the visitors are the drivers. As in the real world, there can be traffic jams, wrong signs, blind alleys, and so on. The content providers, as the road workers, need information about their users to make possible Web site adjustments. However, the content providers have a big advantage in comparison with the road workers – Web logs. Web logs store every motion on the provider's Web site. So the providers need only a tool to analyze these logs. This tool is called Web Usage Mining.

Web Usage Mining is a part of Web Mining – knowledge discovery techniques focused on Web analysis. Web Usage Mining methods analyze moves on Web sites and give the content providers welcomed feedback. Contrary to the simple statistics, which is fine for small Web sites, they provide meaningful insight into how a Web site is being used. There are various application domains like web site design, business and marketing decisions support, personalization, usability studies, etc. For example, e-shop providers can, on the basis of the analysis, alter the recommended merchandise or restructure the Web site to be more user friendly.

The Web structure is also crucial for information systems which are being used by thousands of users. The analysis of such an information system is the topic of this thesis. The Information System (IS) at Masaryk University is an authenticated Web information system which supports almost a full range of administrative and information functions in the study area.

Our goal is to analyze the IS logs. Chapter Two sets up the basic term definitions and taxonomy used in Web mining. Chapter Three gives a detailed overview of the Web Usage Mining, its methods, techniques, and application areas. Chapter Four introduces Information system at Masaryk University.

Chapter Five presents our analysis on is.muni.cz step by step according to the framework presented in chapter Three. That means identifying data source, data cleaning, preprocessing, pattern discovery, and finally pattern analysis. The last chapter summarizes the results and provides a conceivable continuation of our work.

Chapter 2

Web Mining

People take advantage of the Internet in various ways. They search for information with search engines, do shopping in e-shops, sell items, advertise, and so on. Of course, they want to do it quickly, easily, and smoothly. The content providers are trying to offer their users such high-quality services. To do that, they need a lot of information concerning their Web sites and the users who visit them. Web Mining is a great tool for this purpose. It is a part of knowledge discovery[1] and can be broadly defined as the application of data mining technologies to huge Web data repositories.

2.1 Web Data

The data for Web Mining can be gathered from different sources – i.e. from the server-side, the client-side, proxy servers, the company's database, etc. The data differs in its origin as well as in its classification. Usually, it is categorized into four groups [24]:

- **Content:** The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, the text and graphics.
- **Structure:** Data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of inter-page structure information are hyper-links connecting one page to another.

[1] The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [13].

- **Usage:** Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses. Typically, the usage data comes from an Extended Common Log Format (ECLF) server log.

- **User Profile:** Data that provides demographic information about the users of the Web site. This includes registration data and the user profile information.

2.2 Term Definitions

The clarification of the terms used in the Web mining process is very important. Therefore, this section contains the definitions of the widely used but also sometimes misconceived terms. The others can be found in the Web Characterization Terminology And Definitions Sheet [17]:

- **Web Core**
  The collection of resources residing on the Internet that can be accessed using any implemented version of HTTP as part of the protocol stack (or its equivalent), either directly or via an intermediary.

- **Web Resource**
  A resource, identified by a URI, that is a member of the Web Core.

- **URI**
  The URI specification [6] defines a Uniform Resource Identifier (URI) as a compact string of characters for identifying an abstract or physical resource.

- **Link**
  A link expresses one or more (explicit or implicit) relationships between two or more resources.

- **Web Server**
  A server that provides access to Web resources and which supplies Web resource manifestations to the requester.

- **User**
  The principal using a client to interactively retrieve and render resources or resource manifestations.

- **Web Site** A collection of interlinked Web pages, including a host page, residing at the same network location. "Interlinked" is understood to mean that any of the Web site's constituent Web pages can be accessed by following a sequence of references beginning at the site's host page; spanning zero, one or more Web pages located at the same site; and ending at the Web page in question.

- **Cookie**
  Data sent by a Web server to a Web client, to be stored locally by the client and sent back to the server on subsequent requests.

- **Web Page**
  A collection of information, consisting of one or more Web resources, intended to be rendered simultaneously, and identified by a single URI. More specifically, a Web page consists of a Web resource with zero, one, or more embedded Web resources intended to be rendered as a single unit, and referred to by the URI of the one Web resource which is not embedded.

- **Page View**
  Visual rendering of a Web page in a specific client environment at a specific point in time.

- **User Session**
  A delimited set of user clicks across one or more Web servers.

- **Episode**
  A subset of related user clicks that occur within a user session.

- **Server Session**
  A collection of user clicks to a single Web server during a user session. Also called a *visit*.

2.3 Taxonomy of Web Mining

The first classification of Web mining was given in [11] where it is divided into *Web Content Mining* and *Web Usage Mining*. Nowadays *Web Structure Mining* is considered the third part of Web Mining [24, 19].

We briefly describe Web Content Mining and Web Structure Mining in this chapter. The in-depth description of Web Usage Mining will be presented in the next chapter.

Web Content Mining

Web Content Mining is the process of extracting knowledge from the content of Web pages. Cooley *et al.* [11] differentiate between *an Agent-based approach* and *a Database approach*.

The first one utilizes special agents that search, filter, and categorize documents. They use various techniques (domain characteristic, user profiles, information retrieval techniques, and others) to organize and interpret the discovered information.

The second one stores the semi-structured information from the Web in databases and analyzes them with standard database querying and data mining techniques.

Web Structure Mining

Web Structure Mining studies the hyper-link structure of the Web. It categorizes Web pages and generates information (e.g. similarity and relationship between Web sites). See [22, 14] for sample projects.

Chapter 3

Web Usage Mining

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data [24]. Like other data mining disciplines, it defines several procedures leading to the discovery of the desired knowledge. The three main steps are shown in Figure 3.1. They are preprocessing, pattern discovery, and pattern analysis.

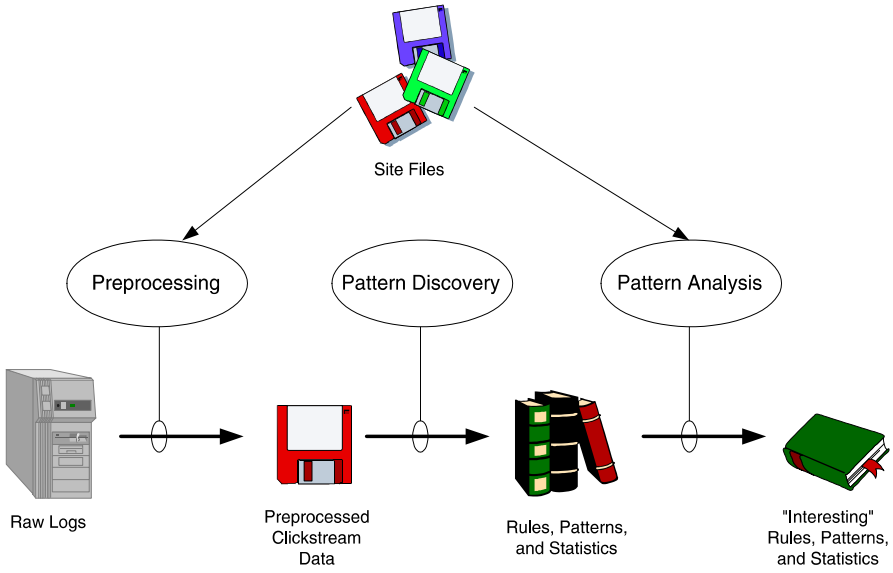

Figure 3.1: The Web Usage Mining process (source [24])

3.1 Data Sources

The data is the basics of a knowledge discovery process. As it is shown in Figure 3.2 there are several possible data sources for the Web Usage Mining process. Each type has its own advantages and a little different focus. For example, server-level data is suitable for mining information from one Web site while client-level logs are optimal for discovering users' behavior during their whole Internet session.

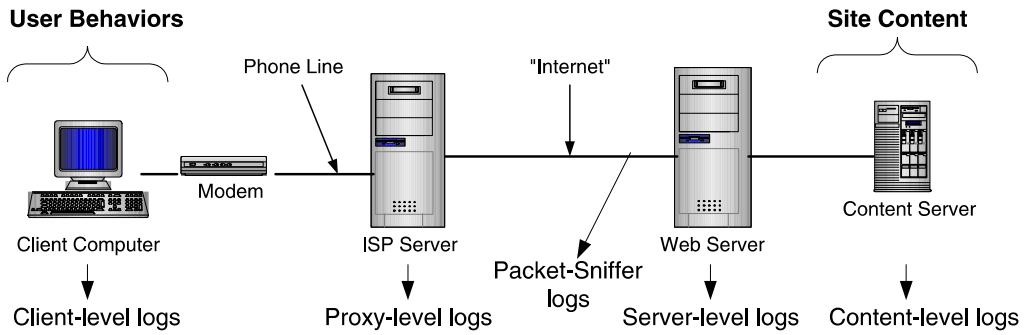

Figure 3.2: Data sources (source [10])

3.1.1 Server-Level Collection

Server-level data collections are often used because of their easy availability – every Web server saves logs by default. The Web server logs are also the best known data source for the Web Usage Mining process.

The log usually contains information about the remote host name or IP address, the user name, time and date, requested URI, server status, and transferred bytes (see The Common Logfile Format [18]). The Extended Log format adds other two entries: referrer and user-agent identification which are very useful for session identification.

From above we can see that server logs store multi-user sessions across a single server. It is helpful for finding out user behavior on one server – the thing that the content providers want.

Unluckily, the Web server logs themselves have serious disadvantages. First, the logs cannot store requests caught by proxies and browser caches. Thus, we have an incomplete picture of user motions and it is very difficult to identify single-user sessions. Second, the logs store only GET parameters of CGI scripts. And third, the stateless of the HTTP protocol causes an ignorance of the page view spent time. Some of these drawbacks can be

reduced with additional tools. For instance, cookies[1] can help with tracking a single user.

Similar work as Web server logs is done by packet sniffers which monitor network traffic by watching server incoming and outgoing TCP/IP packets. Their advantages over Web server logs are that they can save POST parameters of CGI scripts and HTTP headers of requests.

Another server-level data source are logs from content (or application) servers which serve contents for dynamic Web sites. Their advantage is that they can provide information about all incoming parameters, internal state variables, and outgoing data.

3.1.2 Client-Level Collection

Client-level data collections are focused on tracking single-user sessions across single or multiple Web sites. Their main disadvantage is that they are based on user cooperation. There are two ways – remote agents (JavaScript or Java applets) and modified browsers.

The first way is aimed at single-user sessions across a single server. It solves the problem with caching and almost with session identifying. The disadvantages are slow loading in the case of Java applets and no information about page view time – we still do not know when the users close the page.

The best results for single-users/multiple-sites are given by special Web browsers that track every user movement. The browser records how much time the user spends on a Web page, if he pushes the back or reload button, and many other valuable variables. But it is hard to persuade users to use such a special browser. One possibility is to offer them some additional benefits for daily use of modified browsers – see NetZero [3] or Spedia [4].

3.1.3 Proxy-Level Collection

The last data source in our overview are proxy server logs. These logs save multi-user/multi-sites communication and are suitable for characterizing the browse behavior of a group of users sharing the same proxy server.

3.2 Preprocessing

In the preprocessing stage we convert raw data from various data sources into the data suitable for pattern analysis. Preprocessing consists of three

[1] Cookies are small tokens stored on the client's side.

categories – structure preprocessing, content preprocessing and usage preprocessing.

3.2.1 Structure Preprocessing

The knowledge of the Web site structure is necessary for page view identification where frames and dynamic pages are the biggest trouble makers. Furthermore, it is useful for applications like personalization or site modification.

3.2.2 Content Preprocessing

Content preprocessing is based on the Web content mining. The preprocessing stage is useful for filtering input/output data for pattern discovery and analysis. For example, a classification algorithm determines the type of page (see [21, 12] for details) and then we perform analysis only on a subset of all pages.

3.2.3 Usage Preprocessing

Usage preprocessing is the most important stage in Web Usage Mining. The outcome of this stage are mineable objects representing the particular Web site. The diagram of the usage preprocessing process is shown in Figure 3.3. The difficulty of each step varies according to the used Web site technologies. For example, mining server sessions from the dynamic Web site could be less difficult than mining from the static Web site because the static pages are usually caught by proxies, and so many of the requests are missing in the log, while dynamic pages are set not to be stored in proxies, and so the majority of requests are present in the log.

Server Session Identification

Server sessions are initial data for Web Usage Mining. Usually, the data source is a Web server log. However, the raw Web server log is insufficient and has to be adjusted.

Formally, a server session (S) is a time ordered set of page views (V) for a single visit to a Web site, along with some meta data (A) [10].

$$\begin{aligned} S &= [A : V_1, \ldots, V_n] \\ V &= (v_i, h_j, t^f, t^l, t^e, \{d_1, \ldots, d_m\}, c) \\ A &= \{a_1, \ldots, a_k\} \end{aligned}$$

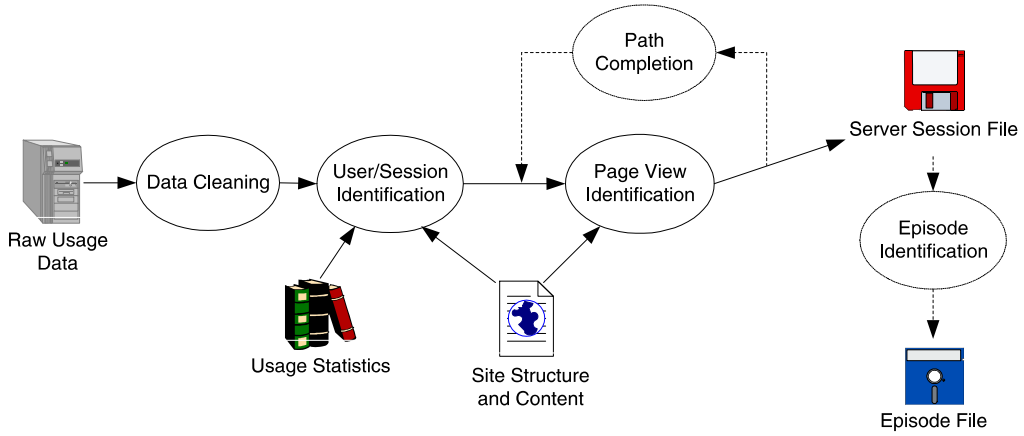

Figure 3.3: Usage preprocessing details (source [10])

Each page view V consists of a view identifier (v_i), the referring page file (h_j), the first request time (t^f), the last request time (t^l), the view end time t^e, an optional set of submitted values ($\{d_1, \ldots, d_m\}$), and the boolean parameter (c) that indicates whether the page view was inferred during the path completion.

The referring page is null if URI was typed, selected from bookmarks, and the like. The first and the last request time corresponds to the page view first and last file request. The first time is used to sort page views in the session. The view end time is usually the first request time of the next request because the real time is unknown[2]. Or it is null if the page view is the last one from the server. The collected meta data A depends on the used technologies. For example, it could be IP address, cookies, user name, etc.

Data Cleaning step puts more data source files into the one data file and filters out the unnecessary records from the logs. Typically, graphics files are useless for mining. Next, records generated by automatic agents have to be removed because they would skew the results. The final task is to normalize URIs. Diverse URIs can lead to the same page view so they have to be the same for the mining process. For example, "`http://is.muni.cz`", "`http://is.muni.cz/`", and "`http://www.is.muni.cz`" lead to the same page view and that is why they should be normalized to "`http://is.muni.cz/`".

[2]This is not true in the case of some client-level collection methods.

User Identification is a very important part of the server session identification process. It is closely linked with user privacy. The more invasive method is used for identifying users, the more accurate results we get. Examples of methods are embed session IDs, cookies, user registration, and so on. If no one is applied, then the single/multiple users with single/multiple IPs problem arises – we do not know how many users are hidden behind one IP address.

Session Identification has one commonly used method when no information about the Web site leaving is available – the thirty minutes timeout [8].

Page View Identification is in most cases important only if frames are used in the mined Web site. We need to know which files are part of one page view and put the relevant records from the data source together. Very helpful in this step is the knowledge of the Web content and structure.

The Path Completion step tries to infer cached pages when request records are not available in the server-level data collections. The utilization of the Web structure and referring pages are basics in this stage.

Episode Identification

Episode identification is an optional step recognizing subsets of user sessions. The goal is to identify users partial interests. For instance, we can define episode as business page views on a news server or shopping cart pages on an e-shop. Besides these manual definitions, there are three general methods based on the assumptions about user browsing behavior.

The page classification to the auxiliary and media pages is based on the assumption that the user browses a Web site until he finds the relevant page – *the media page*. The path to this media page is assembled from *the auxiliary* (or navigational) *pages*.

Generally, there are two types of episodes – *Auxiliary-Media* and *Media-only* (see Figure 3.4). The first one records the path to the media page including the media page whereas the second one records only the media pages. The type of selection depends on the objective of the analysis.

The Page Type Method relies on the page view pre-classification for usage type. Each page view is declared as the auxiliary one or the media one. The disadvantage is that the declaration is constant with time and for

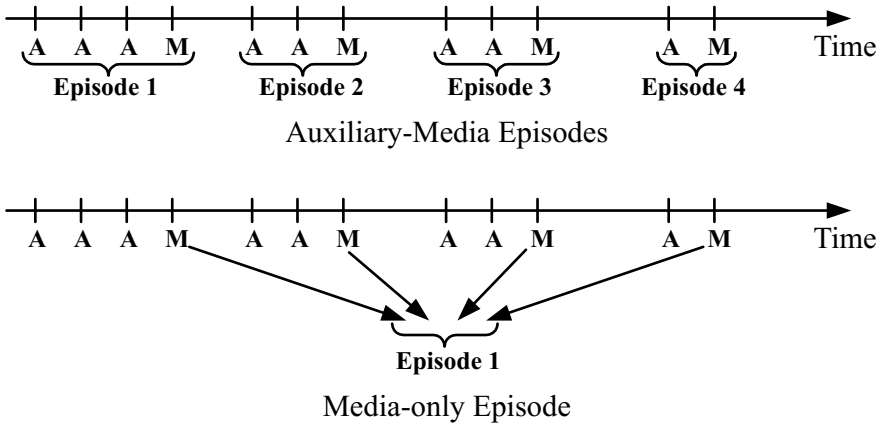

Figure 3.4: Episode types (source [10])

all users. But users can use some pages first as auxiliary and then as the media pages.

The Reference Length Method uses the assumption that the amount of time, which the user spends on the page, relates to whether the page should be classified as the auxiliary one or the media one. It is a supervised method – it calculates the time length cut-off from the percentage estimate of auxiliary pages. See details in [10].

$$t = \frac{-ln(1-\gamma)}{\lambda}$$

where t is time length cut-off,
γ is percentage estimate of auxiliary pages,
λ is reciprocal of observed mean reference length

The last page in a session is considered the media page and it is not counted in the mean. The sample graph for the IS is shown in Figure 3.5. The shape of the curve is similar to the exponential distribution from which the above formula for the cut-off computation is inferred.

Maximal Forward Reference Method The episode is defined as a set of pages in the path from the first page in a user session up to the page before a backward reference is made. The backward reference is a page that is already contained in a set of pages for the current episode; the forward reference is a page that is not yet contained in this set. A new episode is started when the next forward reference is made. The first page of every episode is the first

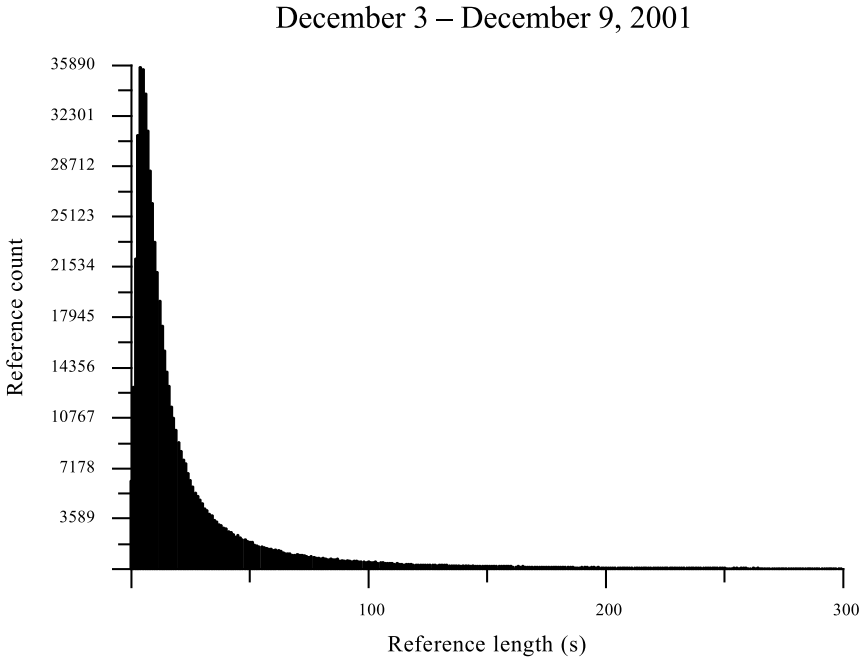

Figure 3.5: Reference lengths

forward reference in a user session. The assumption is that maximal forward references are the media pages and the other pages are the auxiliary pages. This approach needs no input parameter that is based on an estimate but its accuracy goes down for more connected Web sites.

Figure 3.6 shows a sample server session. For example, the forward references are numbers 1, 2, 3, 6; the backward references are numbers 4, 5, 9, 11; episodes are A-B-C-D, A-B-E-G-H; media pages are D, H, W; and auxiliary pages are A, B, C, E.

3.3 Pattern Discovery

Discovering patterns on a Web site is underlying for understanding how users use this Web site. There are several techniques based on statistics, machine learning, pattern recognition, etc.

3.3.1 Statistical Analysis

Statistical analysis is the most common method. We can compute various kinds of descriptive statistics measurements like frequency, mean, median,

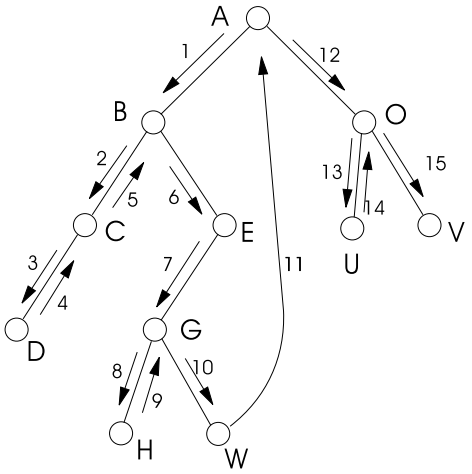

Figure 3.6: Sample traversal patterns (source [9])

and so forth on various values like page views, viewing time, or length of the navigation path.

Although the statistical analysis is not detailed, it is useful for improving system performance, enhancing system security, or facilitating site modification. For example, we can detect unauthorized entry points to our Web site.

3.3.2 Frequent Itemsets and Association Rules

Frequent itemsets are sets of pages which are visited frequently together in a single server session. Only the list of session IDs and URLs is used during this process (the URL order and number of occurrences in the session are irrelevant). Support is often utilized to limit the number of discovered patterns. Support of the subset $\{i_1, \ldots, i_n\}$ from a set D is defined as

$$S(i_1, \ldots, i_n) = \frac{count(\{i_1, \ldots, i_n\} \in D)}{count(D)}$$

After the frequent itemsets discovering, we can calculate for each itemset the *interest* to objectively rank them. Interest is defined as the support of the frequent itemset divided by the probability of all of the items appearing together in a set if the items are randomly and independently distributed [10].

$$I(i_1, \ldots, i_n) = \frac{S(i_1, \ldots, i_n)}{\prod_{j=1}^{n} S(i_j)}$$

Any set of n frequent items can be broken into n separate *association rules*. For example, from the frequent itemset `A` and `B`, we gain two association rules: `A` $\Rightarrow$ `B` and `B` $\Rightarrow$ `A`. The *confidence* of an association rule is the fraction of sessions where the subsequent and the antecedent are present and sessions where only the subsequent is present. For the rule $i_a \Rightarrow i_{s1}, \ldots, i_{sn}$ it is

$$C(i_a \Rightarrow i_{s1}, \ldots, i_{sn}) = \frac{S(i_a, i_{s1}, \ldots, i_{sn})}{S(i_a)}$$

Frequent itemsets and association rules application areas are: business intelligence (e.g. cross promotional opportunities), Web site restructuring, and documents pre-fetching.

3.3.3 Clustering

Clustering is a technique to group together a set of items having similar characteristics. There are two most common cluster types in the Web Usage domain.

- **User clusters**
  The goal is to create groups of users which have common browsing patterns. For instance, this knowledge is useful for personalizing Web content for users: Users in cluster "2" access pages about sports and games.

- **Page clusters**
  Page clusters group together pages with comparable content. For example, these clusters are useful for search engines – they can create a page with links to similar sources: Pages about Doom, Heretic, and Quake belong to the same usage cluster.

3.3.4 Classification

In this task we map data items into one or several predefined classes. The most frequent is developing a profile of users belonging to a particular class. The popular classifying methods are decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, and Support Vector Machines.

3.3.5 Sequential Patterns

This technique tries to find inter-session patterns like one set of items is followed by another set of items. For example, The Quake page is accessed

after the Doom page in 60 percent of time. This knowledge is good for the future visit pattern prediction and hence for placing advertisement aimed at a certain user group. Another application is trend analysis: The number of the Doom page visitors was increasing until February and since then it has been decreasing.

3.3.6 Dependency Modeling

The goal is to establish a model capable of representing significant dependencies among various variables concerning the Web domain. The used techniques are Hidden Markov Models and Bayesian Belief Networks. We can, for example, create a model describing a user behavior on the IS – from the first time visitor to the regular user. Thanks to this knowledge we can improve the documentation and on-line help. It could be good for the predicting of the future Web resource consumption, too.

3.4 Pattern Analysis

Pattern analysis is the final step in the data mining process. It turns the discovered patterns, rules, and statistics into knowledge. The knowledge, we are searching for, has to be somehow interesting for us. The differentiation, between what is interesting and what is uninteresting, is a subjective topic. One can be interested in the most frequent patterns (e.g. marketing analyst), whereas the other is interested in the unusual patterns which have low frequency (e.g. security analyst).

3.4.1 Interestingness Measures

Interestingness measures are classified into four dimensions; a detailed survey is presented in [15]. The dimensions are *representation*[3] (type of patterns measure is applicable to, e.g. classification rules, association rules, etc.), *foundation*[4] (nature of the methodology, e.g. probabilistic, distance, etc.), *scope* (single pattern or rule set), and *class* (objective or subjective). *Objective* measures rate rules according to the structure of the discovered patterns. *Subjective* measures are based upon the user's beliefs or biases.

Although the measures provide automatic division into interesting and uninteresting patterns, the main task rests with the analyst – he has to

[3]also known as *pattern-form*
[4]also known as *representation*

choose the proper measure and set the threshold to get the interesting results he wants.

3.4.2 Tools

Any tool or filter which processes the data obtained from the pattern discovery step can be generally called a pattern analysis tool. The most common are query tools because they allow an analyst to filter and sort the results as he likes. More advanced are On-line Analytical Processing (OLAP) tools which give analyst a multidimensional view of the data. An important role is played also by visualization tools presenting the interesting results in graphs and charts.

3.5 Applications

Figure 3.7 shows major application areas in Web Usage Mining and software applications for these areas. WebSIFT [12] and Web Utilization Miner [23] are general mining tools – they are not focused on any particular subcategory.

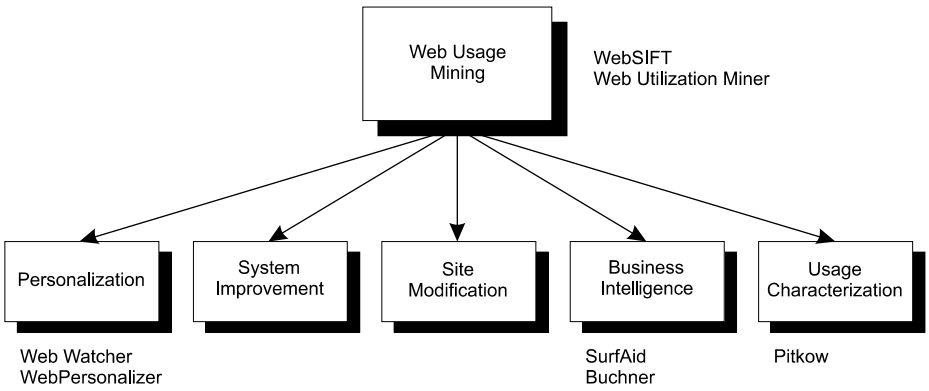

Figure 3.7: Major application areas in Web Usage Mining

3.5.1 Personalization

This is the most popular application of Web Usage Mining. Users like when Web sites fit them. Hence, many Web sites utilize user profiles, dynamic recommendations, individualized marketing, etc. The software application representatives are, for example, WebPersonalizer [20] and WebWatcher [16].

3.5.2 System Improvement

Service quality, security, and performance are important attributes of each Web service. Web Usage Mining prepares groundwork for developing policies for Web caching, load balancing, data distribution, or intrusion, fraud, and break-ins detection.

3.5.3 Site Modification

The user satisfaction is an important criterion for Web sites. For instance, it is crucial to e-commerce. Web Usage Mining provides detailed feedback for site designers who can then adapt the site right to the user's needs.

3.5.4 Business Intelligence

The information about customer behavior in e-shops is key data for marketers. Projects as [7, 5] discover marketing and business intelligence from Web data.

3.5.5 Web Usage Characterization

Pitkow *et al.* [8] studied the user browsing behavior with the special Web browser Xmosaic. They obtained detailed information about browsing strategies, statistics about clicking back/forward buttons, saving Web pages, using bookmarks, and so on.

Chapter 4

Information system is.muni.cz

4.1 The goal

The Information System (IS) [2] at Masaryk University is an authenticated Web information system which supports almost a full range of administrative and information functions in the study area. The goal of the IS is to provide academic community with all information it needs for study and work in one place.

4.2 Inside the IS

The IS has been developed for the Web environment because of its widespread accessibility. It has minimum software requirements – any Web browser. There is no need for JavaScript or cookies. But if JavaScript is in use, it makes applications more comfortable. The system is divided into public and authenticated parts. The security is crucial. All requests are logged into a database and all authenticated transactions are made through the secure HTTPS connection.

Every member of the university has been assigned a unique number – UČO (university personal number). The sophisticated and flexible system of access rights guarantees that every user can manipulate only with data belonging to his authority. Every user has its own email address in format "uco@mail.muni.cz" to be reachable by the system or an university member.

The system has been developed by the university team. The main advantages of this approach are low costs and applications tailored to the university needs. It has been developed continuously. Several new applications are introduced each year. For example, entrance examination agenda and student survey were added last year [1].

The IS is developed above the Apache Web server and the Oracle relational database version 8 with Perl5 as the scripting language (there is the mod_perl module in the Apache Web server). The Oracle database holds data for the IS applications as well as for the Web server authentication.

The IS database server runs on Sun Microsystems Sun Fire V880 with four 750 MHz UltraSparc III type processors, 8 GB RAM and six 32 GB hard-disks connected by Fibre Channel technology. The application server is a cluster of six PCs AMD Athlon XP 2400+ with 1 GB RAM. The operating system is Solaris for the database server and Linux for the application servers.

The statistics from the year 2002 [1] shows that the IS registers 75,254 personal records, 43,851 active users (users who can log in), and 26,110 users who have logged into the system sometimes in the year 2002. There were 53,964,276 accesses, a maximum of 423,786 accesses in one day, and a maximum of 9,849 distinct users in one day.

4.3 Applications

In the following list the main domains of applications in the IS are enumerated:

- Catalog Of Course Offerings
- Course Preregistration, Registrations, Seminar Groups, Timetable
- Grading, Exams
- Administration Of Students Records
- Diary For the Teachers
- Entrance Examination Agenda
- Web-based Email System
- Document Database, Publication Database, Database Of Curricula Vitae

Chapter 5

Mining on is.muni.cz

This chapter describes applications of the Web Usage Mining methods and procedures to the data from the Information System of Masaryk University. We follow the standard data mining procedures step by step as they are described in Chapter 3.

5.1 Data Source

As mentioned in the previous chapter, the IS saves information about every request into a database. There is a special mechanism securing the uniqueness of each request. For instance, the IS must not accomplish two identical requests just due to the pushing reload button. All outgoing pages are also set not to be stored in caches. In other words, the database contains the accurate picture of using the IS.

The data available for mining is parallel to the Web server logs. The information about requests is stored in two tables in the Oracle database. The first table `W_LOG` saves general information about requests – user ID (UČO), his IP address, browser agent, requested URL, and date and time of the request. `NULL` value in the `PEOPLE_ID` column means a request by non-authenticated user. All other columns have always some value. The parameters passed to the called scripts are stored in the second table `W_LOG_PARAM`. See the description of all columns in both tables in Table 5.1 and Table 5.2. Sample records from both tables are shown in Tables 5.3 and 5.4.

The analysis was performed on the data between October 31, 2001 and July 15, 2002. This data covers approximately one semester at Masaryk University. There are 32,316,249 records in the log in the selected time period.

Column	Type	Description
ID	NUMBER(38)	request ID
PEOPLE_ID	NUMBER(38)	user's UČO
URL	VARCHAR2(64)	requested URL
ZADANO	DATE	date and time of request
ADRESA	VARCHAR2(64)	user's IP address
BROWSER	VARCHAR2(64)	user's browser agent

Table 5.1: The W_LOG table structure

Column	Type	Description
LOG_ID	NUMBER(38)	request ID
NAZEV	VARCHAR2(32)	variable's name
HODNOTA	VARCHAR2(64)	variable's value

Table 5.2: The W_LOG_PARAM table structure

5.2 Usage Preprocessing

The raw log from the IS is not proper for the mining process. That is why we first apply preprocessing tasks to this log as it is described in Chapter 3.2.

5.2.1 Server Session Identification

Data Cleaning

We created two new tables called ISLOG and URL. Only these two tables are used in the mining process due to the disk-space issues. In the first table we store the adjusted requests from the IS log. The second table stores all normalized distinct URLs from the IS log. The structure of tables is shown in Table 5.5 and 5.6.

The ISLOG table is quite similar to the W_LOG table. There are not columns ADRESA and BROWSER because we do not need them. We focused only on the analysis of the authenticated users and these columns would be useful if we analyzed all users – then we would need this data for the user identification. There is an extra column DURATION storing how much time a user spent on the requested page. The page identification is stored in the column URL –

ID ADRESA	PEOPLE_ID	URL BROWSER	ZADANO
51801114 ph3-7c53-dial043.sbone.cz	66283	https://is.muni.cz/auth/predmety/predmet.pl Mozilla/4.0 (compatible; MSIE 5.12; Mac_PowerPC)	2002-01-03 00:00:08
51801116 brnoc-197.dialup.vol.cz	50704	https://is.muni.cz/auth/student/zapis.pl Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt)	2002-01-03 00:00:08
51801118 ppp242.uo.worldonline.cz	53067	https://is.muni.cz/auth/mail/ Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt)	2002-01-03 00:00:10
51801119 ppp211.ostrava.worldonline.cz	14329	https://is.muni.cz/auth/rozvrh/rozvrh_zobrazeni.pl Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt)	2002-01-03 00:00:10

Table 5.3: Sample W_LOG table rows

LOG_ID	NAZEV	HODNOTA
51801114	studium	90787
51801114	fakulta	1441
51801114	obdobi	1843
51801114	kod	VVVT5_PCG2
51801114	zpet	https://is.muni.cz/auth/student/zapis.pl?fakulta=1441;obdobi=184
51801116	fakulta	1433
51801116	studium	50097
51801116	obdobi	-
51801118	show	one
51801118	slozka	50267
51801118	select-mail	2662077
51801119	obdobi	1783
51801119	save_obd_pref	1
51801119	fakulta	1422
51801119	studium	11835

Table 5.4: Sample W_LOG_PARAM table rows

Column	Type	Description
ID	NUMBER(38)	request ID – the same as in the W_LOG table
UCO	NUMBER(10)	user ID (UČO)
URL	NUMBER(5)	requested page ID (foreign key into the URL table, column ID)
DATETIME	DATE	date and time of request
DURATION	NUMBER(10)	how long did the user spend on the page

Table 5.5: The ISLOG table structure

Column	Type	Description
ID	NUMBER(5)	URL ID
URL_STR	VARCHAR2(64)	normalized URL string

Table 5.6: The URL table structure

the number is a foreign key to the URL table where the string representation of the requested page is stored. This data organization consumes less space and it is more appropriate for the session and episode identification.

Not all records were copied to the ISLOG table. We did not copy the requests of non-authenticated users because we want to know how the IS is used by university members and not by accidental visitors. The positive result of this decision is that later we will have no problems with the user identification. Next, we did not copy image URLs (these URLs contain "/auth/design/img" text).

Each URL string was normalized before insertion into the URL table. We trimmed off the part identifying the server because the IS server has several aliases – for example, is.muni.cz, www.is.muni.cz, and the like. We added the default script name "index.pl" if the URL had not ended with the script name. Finally, we corrected the URLs showing an attachment of e-mails. The last part of such URLs is the filename of an attached file and we cut off this filename part because it is not valuable for analysis. In fact, it would make noise in the data.

The W_LOG_PARAM table is ignored – we do not consider any script parameters during the analysis. Of course, we lost certain information. For example, we cannot look for the incorrect parameter values and the sequence how a user reached the request with the incorrect value. These sequences

could be helpful for security analysis. During our variable examination we found many requests with doubtful values. Several of them are shown in Table 5.7. The value (column HODNOTA) of beh should be a number, the value of m_nar_text should be empty or a town name, and the value of cislo_op should be a number or a number with two letters.

LOG_ID	NAZEV	HODNOTA
52255704	beh	1www.magick.cz
55029213	m_nar_text	27.8.1983
59931151	cislo_op	ph Brno-stred596475

Table 5.7: Doubtful parameters example

The consequence of this step was reduction in the number of items in the log from 32,316,249 to 19,253,857. There were 9,059,854 accesses of the non-authenticated users. The rest, 4,002,538 accesses, were images. We recognized 467 scripts.

User Identification

Thanks to the log filtering of only the authenticated users we have this step for free. Every row in the database includes the user ID in the column UCO.

Session Identification

We store information about server sessions in two tables – SEQ and SES. The first table stores information about unique server sessions in the given time period. The common thirty-minute time-out was used as the threshold. The second table assigns users to these unique server sessions. The structure is shown in Tables 5.8 and 5.9. The type of the sequence is used for the sequence differentiation in different time periods. The number how many times each user accomplished the unique server session is saved in the column POCET of the SES table.

The first issue, which had to be solved, was the boundary of the time periods. We cannot just start and end creating sessions at the midnight because then we would get incomplete sessions. So, when the session spreads over the threshold time at the boundary of the time period – it means within the first (last) thirty minutes of the first (last) day – we search the previous (next) day, too, until we find two requests with the thirty-minute interval.

Column	Type	Description
ID	NUMBER(38)	sequence ID
TYP	NUMBER(3)	sequence type
ORD	NUMBER(5)	the placings of the URL in sequence
URL	NUMBER(5)	requested URL

Table 5.8: The SEQ table structure

Column	Type	Description
ID	NUMBER(38)	sequence ID
UCO	NUMBER(10)	user ID (UČO)
POCET	NUMBER(10)	the frequency of sequence and user

Table 5.9: The SES table structure

During the sequence creation we came across a problem with repetitive URLs. What does it mean when the same URL repeats in series in sequences? We accept the hypothesis that there is something interesting on that page view and that is why the user accesses this view multiple times. Hence, we merged these URLs into one and count up the duration times to raise the probability that it is a media page.

The examined part of the IS log was divided into many parts according to the different time periods. We created one sequence type for each week. Next, we created three main sequence types – registration period (requests between December 3, 2001 and March 17, 2002), teaching period (February 25, 2002 — May 31, 2002), and exam period (May 27, 2002 — July 12, 2002). The boundaries of the selected time periods were chosen to be the best representation of each period for the majority of faculties at Masaryk University. The registration period includes the exam period of the fall semester so the statistics are affected by this fact. Finally, the sequence type covering the whole semester (December 3, 2001 — July 12, 2002) was created.

Page View Identification

We do not proceed the page view identification step because the IS does not use frames. However, there are scripts which call other scripts. For example, the script `/auth/ucitel/seznam_foto.pl`, which shows the list of students with their photos, calls `/auth/lide/foto.pl`, the script showing the user's photo. The called script appears in the log behind the source script for many times – as many as is the number of shown people. Thanks to the script merging, the called script appears only once in the sequences.

During the pattern analysis we have to take into consideration such script pairs.

Path Completion

We do not perform this step because there are all accesses to the IS in the IS log.

5.2.2 Episode Identification

We performed the episode identification step to discover the media pages. For this purpose we utilized the reference length method which is described on page 13. The reference length mean value was calculated for every week and its value is approximately 50 seconds. We created two episode types. One with 75 percent of auxiliary pages and the other with 85 percent. The first one corresponds to the idea that users find the media page after three auxiliary pages; in the second case it is after five or six pages. Thus, the time length cut-offs are

$$t_1 = \frac{-ln(1-0.75)}{1/50} \doteq 70$$

$$t_2 = \frac{-ln(1-0.85)}{1/50} \doteq 95$$

For the media pages discovering we use the same framework as for the server sessions discovering. The only difference is the threshold – in this case we set it to 70 or 95 seconds (instead of 30 minutes).

5.3 Pattern Discovery and Analysis

The tools we used during the pattern analysis and discovery stage are the SQL of the Oracle database and the Python programming language for the query output formatting.

5.3.1 Simple Statistical Analysis Of the Time Periods

At first, we computed a few statistical values over created server sessions. They are shown in Table 5.10. SL means session length (in pages). In fact, the number of displayed pages in each session is bigger than the session length due to page merging[1].

[1]See the reason for page merging on page 26.

Start date	End date	Number of sessions	Number of users	SL max.	SL avg.	SL 0.5 perc.
Dec 03, 2001	Dec 09, 2001	41,956	11,365	668	15.53	11
Dec 10, 2001	Dec 16, 2001	43,797	11,808	365	14.72	11
Dec 17, 2001	Dec 23, 2001	43,487	11,772	330	14.15	10
Dec 24, 2001	Dec 30, 2001	12,081	6,062	305	13.28	9
Dec 31, 2001	Jan 06, 2002	35,377	11,699	251	14.23	11
Jan 07, 2002	Jan 13, 2002	56,652	13,481	452	14.04	10
Jan 14, 2002	Jan 20, 2002	51,687	12,679	397	13.11	10
Jan 21, 2002	Jan 27, 2002	45,750	12,420	471	14.01	10
Jan 28, 2002	Feb 03, 2002	46,468	13,025	708	14.63	11
Feb 04, 2002	Feb 10, 2002	42,809	12,033	415	14.22	10
Feb 11, 2002	Feb 17, 2002	46,858	12,373	579	14.98	11
Feb 18, 2002	Feb 24, 2002	43,549	12,813	583	13.68	10
Feb 25, 2002	Mar 03, 2002	46,929	12,915	662	13.26	10
Mar 04, 2002	Mar 10, 2002	38,341	12,161	622	13.64	10
Mar 11, 2002	Mar 17, 2002	33,731	11,077	902	13.90	10
Mar 18, 2002	Mar 24, 2002	32,789	10,363	963	14.05	10
Mar 25, 2002	Mar 31, 2002	29,246	9,737	1,329	13.78	10
Apr 01, 2002	Apr 07, 2002	27,283	9,442	1,281	13.45	9
Apr 08, 2002	Apr 14, 2002	33,073	10,060	625	13.46	10
Apr 15, 2002	Apr 21, 2002	35,900	10,189	1,071	13.69	10
Apr 22, 2002	Apr 28, 2002	37,033	10,481	615	12.53	9
Apr 29, 2002	May 05, 2002	35,771	10,411	447	13.29	10
May 06, 2002	May 12, 2002	37,477	10,869	1,108	13.22	9
May 13, 2002	May 19, 2002	46,213	11,587	1,593	13.25	10
May 20, 2002	May 26, 2002	57,667	12,217	1,776	13.06	10
May 27, 2002	Jun 02, 2002	55,718	12,151	472	13.23	10
Jun 03, 2002	Jun 09, 2002	56,977	12,066	455	13.26	10
Jun 10, 2002	Jun 16, 2002	48,429	11,673	847	12.84	10
Jun 17, 2002	Jun 23, 2002	48,944	11,630	535	12.54	9
Jun 24, 2002	Jun 30, 2002	45,237	11,463	454	13.28	10
Jul 01, 2002	Jul 07, 2002	23,000	8,499	381	12.88	9
Jul 08, 2002	Jul 14, 2002	19,955	8,155	456	14.56	10
Dec 03, 2001	Mar 17, 2002	626,960	18,857	902	16.01	12
Feb 25, 2002	May 31, 2002	539,299	18,142	1,776	15.27	11
May 27, 2002	Jul 12, 2002	294,915	16,495	847	14.59	11
Dec 03, 2001	Jul 12, 2002	1,291,386	19,645	1,776	16.20	12

Table 5.10: The IS server sessions statistics

The last row in Table 5.10 shows statistics for the whole semester. 1,291,386 sessions were accomplished by 19,645 users. The distribution of the session lengths is similar to the distribution of the reference lengths except it is more steeper. Ninety percent of session lengths is shorter than twenty pages and fifty percent of the session lengths is shorter or equal to five pages. The maximum session length of 1,776 pages is a big extreme.

There are only 24 sessions longer than 500 pages.

The in-depth analysis of these sessions showed interesting results. Approximately three quarters of these sessions were accomplished by students and consisted almost only of the page view displaying the university member photo. The photos were probably automatically downloaded by some script because there are small reference lengths in the log. If so, the IS photo policy was violated. But a more detailed study should be performed for confirmation. The other long sessions were accomplished by study department officers during registration and exam periods.

Shorter session analyses show that the ratio is decreasing – only 41 percent of sessions contain a majority of photo page views from sessions longer than 100 pages. The utilizing of this knowledge could lead to building an application which automatically discovers suspicious sessions, shows the analyst detailed information about the visited pages in these sessions, reference lengths, etc., and so gives him sufficient data for decision.

Session length averages and 0.5 percentiles show the common usage of the IS. The values support a hypothesis that most users come to the IS to do a couple of things (to read mails, subscribe exam term, check/write exam results, etc.) and then leave the system – the usual session length is around 10 pages. Because of the session lengths distribution, the 0.5 percentile has bigger significance than the average.

Figure 5.1 shows the graph of the IS usage during each week of the spring semester 2002. The points on the x-axis mark weeks; the first day of each week is used as the description. The solid black line draws the number of sessions, the dotted black line draws the number of users, and the solid gray line draws the ratio between these two values. The ratio line shows the intensity of using IS – it varies from two sessions per one user in one week to 4.7 sessions.

The graph shows that there is a significant activity decrease during Christmas and summer holidays. Another one is at the beginning of April. On the contrary, the access maximums are made during exam periods (January and February, end of May and June). The maximum before the spring exam period is probably made by students registering themselves for the exam terms. This data is not surprising – the main applications for students (who are the biggest group of users) are exam and registration-enrollment agenda.

5.3.2 In-depth Script Analysis

The next task, we performed after simple statistical analysis, is a more detailed analysis of several scripts. We aimed at the path analysis, e.g., how users get to the script, where they continue, and the like.

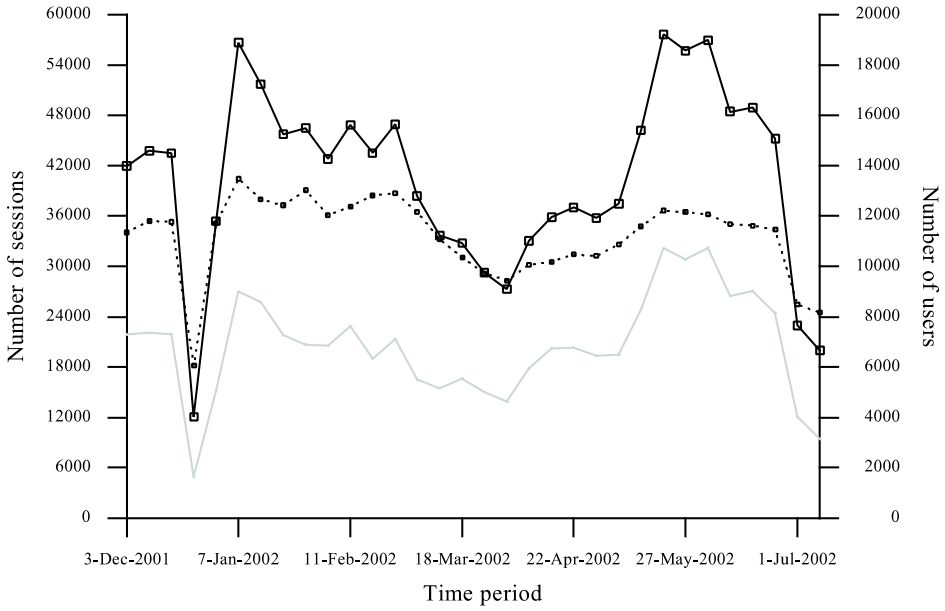

Figure 5.1: Session and user statistics

Script Selection

The scripts for in-depth analysis were chosen on the basis of their usage. We examined the media pages found during the episode identification step. From both sets[2] we filtered out the pages with the lower occurrence than one percent of all media pages and which were visited by less than one percent of distinct users. The results are shown in Table 5.11 (episodes with the threshold 70 seconds – Ep70) and Table 5.12 (episodes with the threshold 95 seconds – Ep95).

The tables contain statistics for scripts in the following time periods: the registration period, the teaching period, the exam period, and the whole semester. For each script and given time period these values are shown: the number of occurrences as a media page, the number of distinct users who accessed the script as a media page, and the ratio between these two values (pages divided by users). Symbol "–" means that the script did not have the sufficient occurrence in this time period. The row "Total" contains the total number of media pages and distinct users in the given time period.

From the Ep70 and Ep95 comparison could be seen that the second one is inappropriate for the last two time periods (the exam period and the whole

[2]We have two different episode types – with 70 and 95 seconds threshold. See page 28 for details.

Script name	Regis. pages	Regis. users	Regis. ratio	Teach. pages	Teach. users	Teach. ratio	Exam pages	Exam users	Exam ratio	Semes. pages	Semes. users	Semes. ratio
/auth/dok/index.pl	43,931	3,722	11.80	60,091	3,934	15.27	13,603	2,357	5.77	99,823	4,733	21.09
/auth/hry/01/index.pl	24,779	1,991	12.45	19,554	1,423	13.74	9,581	798	12.01	48,304	2,922	16.53
/auth/index.pl	245,268	17,237	14.23	208,718	15,930	13.10	90,832	13,181	6.89	478,645	18,405	26.01
/auth/lide/foto.pl	48,588	7,693	6.32	46,111	7,264	6.35	20,077	4,898	4.10	100,058	10,145	9.86
/auth/lide/index.pl	42,187	7,300	5.78	37,346	6,813	5.48	13,450	4,242	3.17	80,383	9,603	8.37
/auth/mail/index.pl	235,191	11,826	19.89	264,630	11,219	23.59	96,451	8,998	10.72	523,590	13,673	38.29
/auth/mail/mail.pl	34,688	7,806	4.44	34,081	7,612	4.48	13,024	4,761	2.74	71,947	10,213	7.04
/auth/mail/mail.pl/	–	–	–	18,918	3,371	5.61	–	–	–	–	–	–
/auth/mail/mail_posli.pl	89,027	7,843	11.35	97,341	7,647	12.73	36,481	5,589	6.53	196,065	10,170	19.28
/auth/predmety/predmet.pl	32,160	7,686	4.18	17,769	5,450	3.26	9,686	3,254	2.98	48,764	9,076	5.37
/auth/rozvrh/rozvrh_zobrazeni.pl	26,633	5,019	5.31	–	–	–	–	–	–	–	–	–
/auth/seminare/student.pl	116,007	7,335	15.82	27,617	4,907	5.63	37,238	2,293	16.24	160,720	7,419	21.66
/auth/student/index.pl	28,462	6,470	4.40	23,714	4,853	4.89	17,232	3,805	4.53	62,799	8,496	7.39
/auth/student/moje_znamky.pl	108,492	12,879	8.42	52,295	10,857	4.82	84,359	10,517	8.02	219,524	14,441	15.20
/auth/student/prihl_na_zkousky.pl	319,509	12,273	26.03	221,164	10,894	20.30	136,671	10,182	13.42	635,285	13,116	48.44
/auth/student/zapis.pl	233,830	14,234	16.43	93,055	10,877	8.56	69,009	8,103	8.52	325,711	14,735	22.10
/auth/ucitel/znamky.pl	50,078	1,243	40.29	28,387	1,134	25.03	34,480	1,183	29.15	99,649	1,447	68.87
Total	2,127,132	18,859		1,621,387	18,144		832,663	16,519		4,007,446	19,669	

Table 5.11: Episode statistics for threshold 70 seconds

Script name	Regis. pages	Regis. users	Regis. ratio	Teach. pages	Teach. users	Teach. ratio	Exam pages	Exam users	Exam ratio	Semes. pages	Semes. users	Semes. ratio
/auth/dok/index.pl	37,479	3,677	10.19	13,703	1,563	8.77	–	–	–	–	–	–
/auth/hry/01/index.pl	23,036	1,942	11.86	8,790	499	17.62	–	–	–	–	–	–
/auth/index.pl	218,179	16,902	12.91	56,251	6,421	8.76	1	1	1.00	9	1	9.00
/auth/lide/foto.pl	41,765	7,113	5.87	21,804	2,947	7.40	–	–	–	–	–	–
/auth/lide/index.pl	32,366	6,702	4.83	11,765	2,532	4.65	–	–	–	–	–	–
/auth/mail/index.pl	196,407	11,349	17.31	53,598	3,746	14.31	–	–	–	–	–	–
/auth/mail/mail.pl	29,268	7,250	4.04	8,116	2,221	3.65	–	–	–	–	–	–
/auth/mail/mail_posli.pl	79,141	7,644	10.35	21,384	2,447	8.74	–	–	–	–	–	–
/auth/predmety/katalog_plneni.pl	–	–	–	10,209	300	34.03	–	–	–	–	–	–
/auth/predmety/predmet.pl	24,286	6,849	3.55	–	–	–	–	–	–	–	–	–
/auth/predmety/sylaby_plneni.pl	–	–	–	–	–	–	–	–	–	5	1	5.00
/auth/rozvrh/rozvrh_zobrazeni.pl	22,648	4,787	4.73	–	–	–	–	–	–	–	–	–
/auth/seminare/student.pl	100,851	7,226	13.96	–	–	–	–	–	–	–	–	–
/auth/student/index.pl	24,674	5,959	4.14	7,301	1,458	5.01	–	–	–	–	–	–
/auth/student/moje_znamky.pl	96,036	12,315	7.80	14,589	3,614	4.04	–	–	–	–	–	–
/auth/student/prihl_na_zkousky.pl	268,422	12,159	22.08	52,742	3,475	15.18	–	–	–	–	–	–
/auth/student/zapis.pl	188,617	14,113	13.36	25,833	3,696	6.99	–	–	–	–	–	–
/auth/ucitel/index.pl	–	–	–	–	–	–	14	1	14.00	33	1	33.00
/auth/ucitel/seznam.pl	–	–	–	6,829	880	7.76	1	1	1.00	2	1	2.00
/auth/ucitel/zkusebni_terminy.pl	–	–	–	7,379	506	14.58	–	–	–	–	–	–
/auth/ucitel/znamky.pl	43,582	1,236	35.26	22,249	1,056	21.07	27	1	27.00	65	1	65.00
Total	1,791,404	18,859		514,486	8,037		43	1		114	1	

Table 5.12: Episode statistics for threshold 95 seconds

semester). The reason is maybe the high threshold or the wrong estimate of the auxiliary pages in sessions. Surprising could be that the Ep95 found more distinct media pages than the Ep70. But only until we realize that the Ep95 contains less media pages and less users so one percent threshold means also less filtering.

From all the scripts found we chose these four ones for a detailed analysis:

- `/auth/dok/index.pl`
- `/auth/predmety/katalog_plneni.pl`
- `/auth/student/prihl_na_zkousky.pl`
- `/auth/ucitel/znamky.pl`

Methodology

The analysis of each script comprises the graph of usage and the path analysis. The graph is the same as for the server sessions – on the x-axis there are weeks, on the y-axis the black solid line marks the number of media pages (how many times the given script was considered as the media page), the black dotted line marks the number of users (how many users considered the given script as the media page), and the gray solid line marks the pages/users ratio. The data for the graph comes from the Ep70.

The path analysis explores pages from which users reached the analyzed script and where they continued. The data comes from the server sessions. We analyzed pages in distance one and five around the given script. The pages with distance five are listed in Appendix A. We also tested the longer distance (10, 15, and 20 pages) around the given script but the results were very similar to the five-page distance. The minimum for the listing of a page is 0.5 percent of occurrences from pages in the given group (e.g. all pages just before the given script). One hundred percent is the number of the given script occurrences for the one-page distance and the quintuple of this amount for the five-page distance. The row with "none" shows the percentage of those being the last or the first page in a server session for the given script in the case of the one-page distance.

There are again statistics displayed for four different time periods – the whole semester, the registration period, the teaching period, and the exam period. The listed pages are sorted by the whole semester column. The first part shows pages before the given script, then is the row with the number of occurrences for the given script, and at the end of the table there are pages after the given script.

The second part of the path analysis is the examination of trails between two pages. We picked out one page from the pages before the given script and made an overview of how users reach the given script from the picked one. We called one trail "the pattern". The overview is in the table which has three parts. The first part is statistics of the number of all sessions, the number of the sessions containing the beginning page, the number of sessions containing the given script, the number of sessions containing one or more patterns, and the number of all patterns. The second part shows the number of patterns in one server session and the third part shows the pattern lengths. The threshold for putting a record into a table was set to three percent from the number of all sessions or from the number of all patterns. Finally, we listed the most popular patterns from the whole semester for the trails ten pages length.

Script `/auth/dok/index.pl`

This script shows lists and contents of documents available in the electronic form at Masaryk University. It is an example of a mixed content page – usually, the lists of documents are the auxiliary pages and the documents themselves are the media pages.

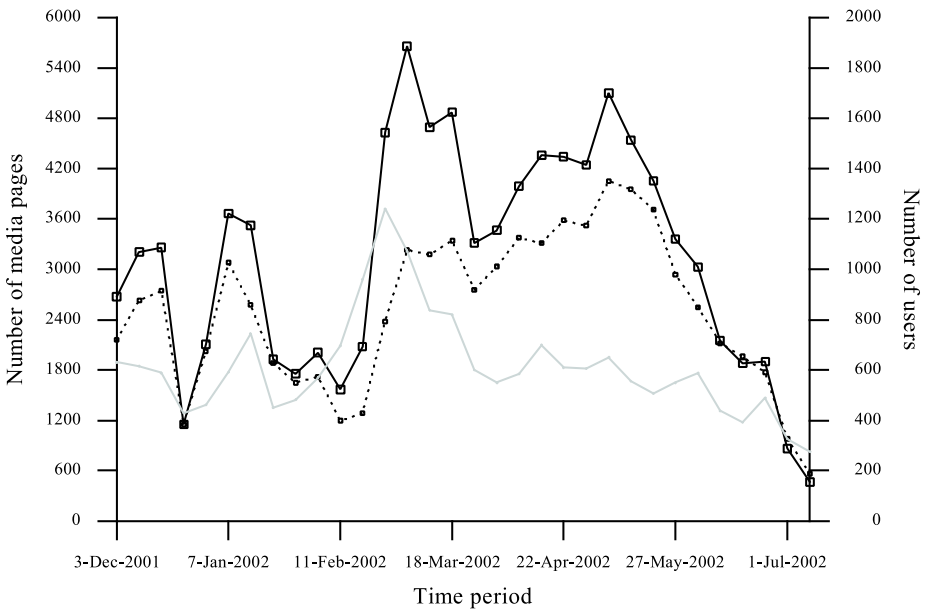

Figure 5.2: The `/auth/dok/index.pl` statistics

The graph shows that the usage maximum is at the beginning of March

and in May. The May maximum is for the showed pages and also for the users. In the March maximum there is the biggest pages/users ratio – each user displayed almost six documents in average. In other time periods the ratio is about three documents per user.

The page of departure for the documents page is in approximately 75 percent of accesses the main page for the authenticated users. In about ten percent the documents page is the first page in a server session. This could be, for example, caused by selecting the link from the bookmarks or following a link from the outside of the IS. The presence of the `/auth/mail/index.pl` script (the IS Web-based e-mail system) and the `/auth/student/prihl_na_zkousky.pl` script (the script for the subscribing to exam terms) is probably caused by a link to a document in a mail or in an exam term description. Almost a half of server sessions containing this documents page ended by this page (around 45 percent). Less (around 43 percent) continued back to the main page. The detailed listings of the preceding and following pages are in Tables 5.13 and A.2.

We do not talk about the user percentage here because the distinct user percentage does not give the accurate numbers to us. The reason is that some users sometimes finish their session on the documents page and sometimes they continue with browsing. For example, in the whole semester, 75 percent of users sometimes continued to the main page and 74 percent of users sometimes finish their session here.

Script name	Semes.	Regis.	Teach.	Exam
/auth/index.pl	75.03%	71.85%	77.75%	74.51%
none	10.06%	11.72%	8.59%	9.74%
/auth/mail/index.pl	2.46%	2.37%	2.86%	2.09%
/auth/student/prihl_na_zkousky.pl	2.27%	2.18%	1.93%	2.36%
/auth/dok/dok_novy.pl	1.51%	1.93%	1.45%	1.27%
/auth/student/moje_znamky.pl	0.93%	1.12%	-	1.96%
/auth/mail/mail.pl	0.87%	0.84%	0.95%	0.69%
/auth/student/zapis.pl	0.58%	0.69%	-	0.77%
/auth/student/index.pl	0.57%	-	0.59%	0.82%
/auth/seminare/student.pl	-	-	-	0.69%
/auth/dok/index.pl	60,458	26,202	34,804	8,947
none	45.17%	43.34%	46.32%	45.20%
/auth/index.pl	43.30%	44.11%	42.96%	42.44%
/auth/dok/dok_novy.pl	1.72%	2.21%	1.62%	1.51%
/auth/mail/mail.pl	1.60%	1.61%	1.61%	1.53%
/auth/mail/index.pl	1.24%	1.10%	1.36%	1.20%
/auth/student/index.pl	0.96%	0.79%	0.86%	1.80%
/auth/student/prihl_na_zkousky.pl	0.83%	0.64%	0.82%	0.80%
/auth/lide/index.pl	0.72%	0.79%	0.75%	0.58%

Table 5.13: The `/auth/dok/index.pl` whence-where pages

For the pattern analysis the script for the subscribing exam terms

(/auth/student/prihl_na_zkousky.pl) was chosen. The pattern table (Table 5.14) shows that approximately one eighth of the users who visited the documents page, visited the page for subscribing exam terms before it. The documents page is the most popular in the teaching period when more than five percent of all sessions contain at least one request for it. Most users accomplished only one pattern per one server session; a few tens of users did two. Interesting is that in the teaching period four distinct users accomplished this pattern nine times in one server session. The length analysis shows that one half of patterns led through one page – the main page. About 12 — 13 percent of trails led directly or through main page and one other page.

	Semester		Registration		Teaching		Exam	
Number of all sessions	1,291,386		626,960		539,299		294,915	
/auth/student/prihl_na_zkousky.pl	421,722	32.66%	204,281	32.58%	139,747	25.91%	108,804	36.89%
/auth/dok/index.pl	48,265	3.74%	20,454	3.26%	27,656	5.13%	7,380	2.50%
Number of sessions with pattern	6,088	0.47%	2,529	0.40%	3,115	0.58%	1,207	0.41%
Number of all patterns	9,887		3,773		5,001		1,744	
1 occurrence	5,832	95.80%	2,429	96.05%	2,970	95.35%	1,172	97.10%
2 occurrences	199	3.27%	75	2.97%	113	3.63%	31	2.57%
Pattern length 2	1,373	13.89%	570	15.11%	673	13.46%	211	12.10%
Pattern length 3	4,992	50.49%	1,771	46.94%	2,683	53.65%	824	47.25%
Pattern length 4	471	4.76%	194	5.14%	226	4.52%	88	5.05%
Pattern length 5	1,209	12.23%	466	12.35%	568	11.36%	273	15.65%
Pattern length 6	400	4.05%	150	3.98%	210	4.20%	72	4.13%
Pattern length 7	540	5.46%	222	5.88%	247	4.94%	106	6.08%

Table 5.14: The /auth/student/prihl_na_zkousky.pl → /auth/dok/index.pl patterns

The most popular patterns in the whole semester are:

- 4,867 occurrences (49.23%)
  1. /auth/student/prihl_na_zkousky.pl
  2. /auth/index.pl
  3. /auth/dok/index.pl

- 1,373 occurrences (13.89%)
  1. /auth/student/prihl_na_zkousky.pl
  2. /auth/dok/index.pl

- 361 occurrences (3.65%)
  1. /auth/student/prihl_na_zkousky.pl
  2. /auth/index.pl
  3. /auth/student/moje_znamky.pl
  4. /auth/index.pl
  5. /auth/dok/index.pl

Script `/auth/predmety/katalog_plneni.pl`

This script allows the subject attributes editing. Attributes are, for example, the type of the ending, the enrollment restriction, the teaching time period, and so on.

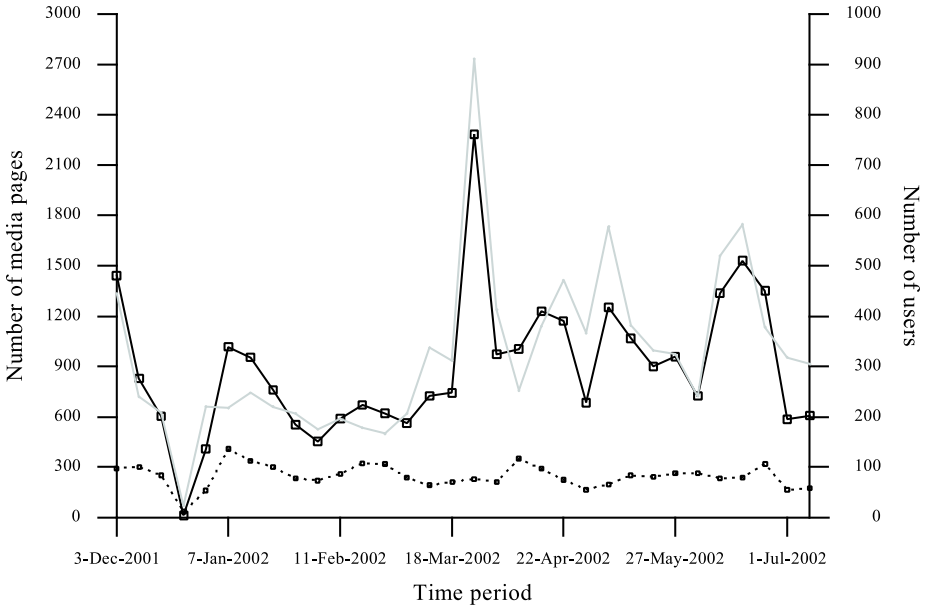

Figure 5.3: The `/auth/predmety/katalog_plneni.pl` statistics

This script was chosen for the analysis because of the high pages/users ratio in the teaching time period of the Ep95 where the ratio was 34.03 pages per user. The high value was confirmed also in the Ep70 – it was almost 30 pages per user in the week between March 25, 2002 and March 31, 2002. This high amount of activity is caused by the study catalog deadline for the next year. The most of accesses was accomplished by teachers from the Faculty of Arts and the Faculty of Medicine. Similarly, a high ratio value in May was accomplished by teachers from the Faculty of Education. The June one is,

on the contrary, achieved by users from various faculties. The graph further shows that the number of users using this application is constant – around 90.

Of course, the departure page for this application is mostly some page concerning the subjects (any page which URI contains "`/auth/predmety/`"; mostly the `/auth/predmety/index.pl`). The length of Table 5.15 points out that users come to and leave this page from/to many pages. It can be also seen in Table 5.15 that the departure page changed during the semester in favor of the most used `/auth/predmety/index.pl`. The other pages usage decreased (e.g., `/auth/predmety/uplny_vypis.pl` or `/auth/predmety/sylaby_plneni.pl`). Interesting is the increase of usage the `/auth/predmety/atributy_plneni.pl` page in the teaching period. As to the following pages, in the registration period few users continued on the `/auth/predmety/atributy_plneni.pl` page. They rather went to the main page (`/auth/index.pl`) or to the `/auth/predmety/sylaby_plneni.pl` page. This turned contrariwise in the teaching and exam periods. The detailed listings of the preceding and following pages are in Tables 5.15 and A.3.

We decided to find out the popular trails from the main page to this application in the pattern analysis part. Approximately one third of the sessions with analyzed script does not contain the main page (see Table 5.16). It means that some users[3] have this page (or a similar one like the main page for subjects – `/auth/predmety/index.pl`) in the bookmarks and they do not need to go through the main page. The number of the multiple pattern occurrences in one session is usually one, to a lesser extent, two or three. The maximum is 17 occurrences accomplished by one user in the teaching period. The length analysis shows that almost one third of patterns led directly, one third led through one page (mostly the `/auth/predmety/index.pl`), and the rest of patterns had longer trails.

The most popular patterns in the whole semester are:

- 2,216 occurrences (31.15%)
  1. /auth/index.pl
  2. /auth/predmety/katalog_plneni.pl

[3]We cannot say one third of users – see page 36 for explanation.

Script name	Semes.	Regis.	Teach.	Exam
/auth/predmety/index.pl	39.80%	36.18%	38.75%	46.61%
/auth/predmety/uplny_vypis.pl	17.18%	19.61%	16.96%	13.54%
/auth/index.pl	14.79%	14.74%	14.57%	14.99%
/auth/predmety/sylaby_plneni.pl	6.90%	9.19%	4.81%	5.71%
none	5.43%	4.84%	5.86%	6.21%
/auth/predmety/katalog_vypis.pl	3.12%	3.06%	4.74%	1.68%
/auth/predmety/atributy_plneni.pl	2.97%	1.14%	5.82%	1.04%
/auth/studium/zapis.pl	1.28%	0.86%	0.91%	2.64%
/auth/ucitel/index.pl	1.05%	1.66%	1.04%	0.75%
/auth/lide/foto.pl	0.89%	0.78%	1.07%	1.07%
/auth/predmety/sablony_plneni.pl	0.71%	0.65%	-	1.10%
/auth/predmety/atributy_predmety_vypis.pl	0.71%	0.91%	0.64%	0.52%
/auth/predmety/predmet.pl	0.57%	1.04%	-	-
/auth/predmety/katalog_hlavni.pl	0.50%	0.54%	-	0.73%
/auth/prezentator/index.pl	-	-	0.76%	-
/auth/predmety/novy_semestr.pl	-	0.59%	-	-
/auth/predmety/katalog_plneni_hromadne.pl	-	-	0.54%	-
/auth/predmety/katalog_plneni.pl	14,979	6,134	6,841	3,448
/auth/predmety/atributy_plneni.pl	16.79%	6.06%	22.16%	22.82%
none	15.72%	16.66%	15.68%	15.46%
/auth/predmety/index.pl	14.89%	14.48%	14.63%	15.43%
/auth/index.pl	14.75%	18.39%	12.70%	14.18%
/auth/predmety/uplny_vypis.pl	12.98%	15.16%	12.15%	10.53%
/auth/predmety/sylaby_plneni.pl	11.11%	14.59%	8.99%	8.03%
/auth/studium/zapis.pl	2.14%	1.50%	1.97%	3.83%
/auth/predmety/katalog_vypis.pl	1.82%	1.86%	2.78%	0.64%
/auth/ucitel/index.pl	1.15%	1.50%	1.02%	1.16%
/auth/predmety/sablony_plneni.pl	0.95%	0.86%	0.56%	1.54%
/auth/predmety/predmet.pl	0.87%	0.93%	1.02%	-
/auth/lide/index.pl	0.50%	0.83%	-	-
/auth/seminare/seminare_plneni.pl	-	0.52%	-	-
/auth/prezentator/index.pl	-	-	0.91%	-
/auth/predmety/osoby_z_pracovist_predmetu.pl	-	0.64%	-	-
/auth/predmety/novy_semestr.pl	-	0.54%	-	-
/auth/lide/foto.pl	-	-	0.51%	0.70%

Table 5.15: The `/auth/predmety/katalog_plneni.pl` whence-where pages

- 1,966 occurrences (27.64%)
  1. /auth/index.pl
  2. /auth/predmety/index.pl
  3. /auth/predmety/katalog_plneni.pl

- 152 occurrences (2.14%)
  1. /auth/index.pl
  2. /auth/predmety/uplny_vypis.pl
  3. /auth/predmety/katalog_plneni.pl

As we can see, the most popular pattern is the direct one although we can see a big percentage number of the `/auth/predmety/index.pl` script as the departure page (see Table 5.15). The reason is that this page precedes

	Semester		Registration		Teaching		Exam	
Number of all sessions	1,291,386		626,960		539,299		294,915	
/auth/index.pl	1,203,489	93.19%	584,292	93.19%	502,899	93.25%	273,561	92.76%
/auth/predmety/katalog_plneni.pl	7,114	0.55%	3,492	0.56%	2,977	0.55%	1,498	0.51%
Number of sessions with pattern	4,621	0.36%	2,406	0.38%	1,935	0.36%	988	0.34%
Number of all patterns	7,113		3,415		2,938		1,559	
1 occurrence	3,697	80.00%	1,984	82.46%	1,557	80.47%	773	78.24%
2 occurrences	666	14.41%	319	13.26%	276	14.26%	141	14.27%
3 occurrences	157	3.40%	69	2.87%	61	3.15%	41	4.15%
Pattern length 2	2,216	31.15%	904	26.47%	997	33.93%	517	33.16%
Pattern length 3	2,427	34.12%	1,192	34.90%	974	33.15%	550	35.28%
Pattern length 4	640	9.00%	317	9.28%	259	8.82%	138	8.85%
Pattern length 5	656	9.22%	352	10.31%	259	8.82%	125	8.02%
Pattern length 6	335	4.71%	185	5.42%	135	4.59%	65	4.17%
Pattern length 7	301	4.23%	182	5.33%	104	3.54%	53	3.40%

Table 5.16: The `/auth/index.pl` → `/auth/predmety/katalog_plneni.pl` patterns

our application in many different trails which have the low percentage of occurrences.

Script `/auth/student/prihl_na_zkousky.pl`

This application helps students with exam planning. They see all subjects they enrolled in the given semester, with the available exam dates and times. The can register and unregister for these terms.

The graph shows that the students register for their exams from approximately one month prior to the exam period. Without considering Christmas vacation in the fall semester, the number of users and displayed pages rises until the second week of January, the second week of the fall exam period. The maximum in the spring semester is reached in the first week of the exam period. The maximum pages/users ratio is reached in the first week of May when the value hits 7.59 pages per user. It is achieved mostly by students of the Faculty of Law and the Faculty of Medicine.

There are no surprising results in the path analysis. The percentage numbers are stable throughout the whole semester. The departure page for this application is either the main index page for the authenticated users or the index page for the students. The next page is in 40 percent of accesses again the main index page. Or the server sessions end by this page at the same probability. About one tenth of accesses continued back to the students index page. The detailed listings of the preceding and following pages are in Tables 5.17 and A.1.

We focused on the relation with page showing the student grades (`/auth/student/moje_znamky.pl`) during the path analysis step. We computed both pattern variants – from the grade page to the exam terms page

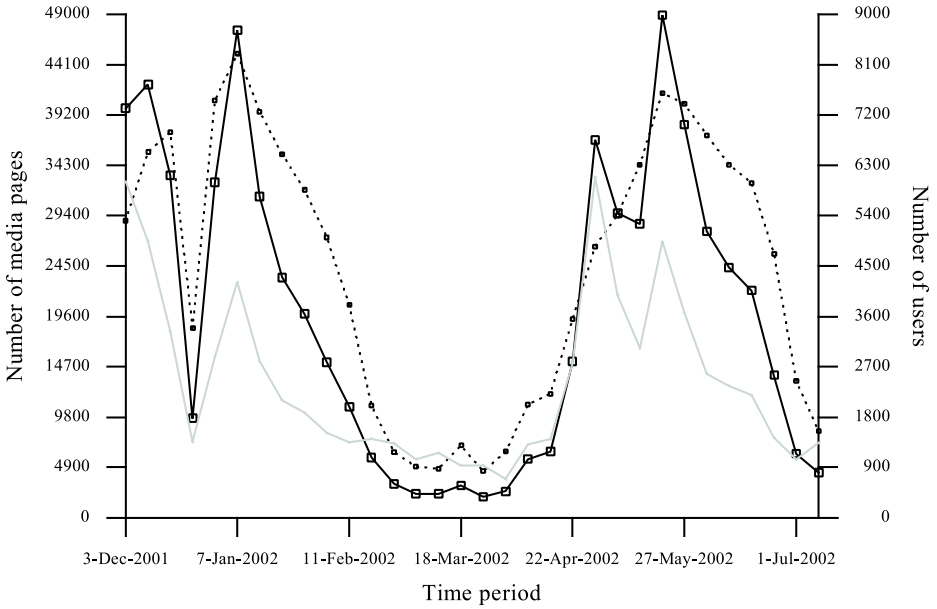

Figure 5.4: The `/auth/student/prihl_na_zkousky.pl` statistics

Script name	Semes.	Regis.	Teach.	Exam
/auth/index.pl	77.53%	77.53%	76.95%	78.01%
/auth/student/index.pl	15.21%	15.61%	15.12%	14.60%
none	2.73%	2.53%	3.25%	2.40%
/auth/student/moje_znamky.pl	1.15%	1.05%	0.86%	1.79%
/auth/mail/index.pl	0.68%	0.61%	0.83%	0.65%
/auth/mail/mail.pl	0.58%	0.56%	0.66%	0.52%
/auth/student/zapis.pl	-	-	-	0.52%
/auth/student/prihl_na_zkousky.pl	532,475	260,053	178,293	132,939
/auth/index.pl	40.33%	40.85%	40.37%	39.13%
none	40.23%	38.83%	41.18%	41.49%
/auth/student/index.pl	11.41%	12.37%	10.07%	11.36%
/auth/mail/mail.pl	1.82%	1.67%	2.09%	1.70%
/auth/mail/index.pl	1.19%	1.10%	1.46%	1.05%
/auth/system/kill.pl	1.14%	1.27%	1.11%	0.88%
/auth/student/moje_znamky.pl	0.87%	0.82%	0.60%	1.42%
/auth/lide/index.pl	0.64%	0.61%	0.70%	0.61%
/auth/student/zapis.pl	-	-	-	0.51%

Table 5.17: The `/auth/student/prihl_na_zkousky.pl` whence-where pages

and vice versa. The numbers of sessions, shown in Tables 5.18 and 5.19, say that the first case is more often – the number of occurrences is about 1,075 to 1,364 higher. This is true except for the teaching period where the second case exceeded the first one by 891 sessions. However, the differences are too small to accept a hypothesis that users first want to know their grades and

	Semester		Registration		Teaching		Exam	
Number of all sessions	1,291,386		626,960		539,299		294,915	
/auth/student/moje_znamky.pl	340,496	26.37%	174,805	27.88%	84,407	15.65%	119,802	40.62%
/auth/student/prihl_na_zkousky.pl	421,722	32.66%	204,281	32.58%	139,747	25.91%	108,804	36.89%
Number of sessions with pattern	43,946	3.40%	25,098	4.00%	10,489	1.94%	14,957	5.07%
Number of all patterns	89,989		46,104		18,912		31,959	
1 occurrence	40,695	92.60%	23,248	92.63%	9,893	94.32%	13,850	92.60%
2 occurrences	2,853	6.49%	1,632	6.50%	530	5.05%	967	6.47%
Pattern length 2	6,099	6.78%	2,722	5.90%	1,527	8.07%	2,380	7.45%
Pattern length 3	68,456	76.07%	34,958	75.82%	14,396	76.12%	24,412	76.39%
Pattern length 4	3,426	3.81%	1,764	3.83%	704	3.72%	1,222	3.82%
Pattern length 5	5,385	5.98%	2,976	6.45%	1,022	5.40%	1,767	5.53%

Table 5.18: The `/auth/student/moje_znamky.pl` → `/auth/student/prihl_na_zkousky.pl` patterns

	Semester		Registration		Teaching		Exam	
Number of all sessions	1,291,386		626,960		539,299		294,915	
/auth/student/prihl_na_zkousky.pl	421,722	32.66%	204,281	32.58%	139,747	25.91%	108,804	36.89%
/auth/student/moje_znamky.pl	340,496	26.37%	174,805	27.88%	84,407	15.65%	119,802	40.62%
Number of sessions with pattern	42,582	3.30%	24,023	3.83%	11,380	2.11%	13,605	4.61%
Number of all patterns	81,876		41,158		19,900		27,479	
1 occurrence	39,901	93.70%	22,536	93.81%	10,811	95.00%	12,727	93.55%
2 occurrences	2,358	5.54%	1,315	5.47%	503	4.42%	768	5.64%
Pattern length 2	4,640	5.67%	2,134	5.18%	1,077	5.41%	1,893	6.89%
Pattern length 3	62,629	76.49%	31,242	75.91%	15,142	76.09%	21,264	77.38%
Pattern length 4	3,408	4.16%	1,724	4.19%	891	4.48%	1,093	3.98%
Pattern length 5	4,751	5.80%	2,512	6.10%	1,224	6.15%	1,383	5.03%

Table 5.19: The `/auth/student/prihl_na_zkousky.pl` → `/auth/student/moje_znamky.pl` patterns

then register for exam terms. Apart from this difference, the both patterns have approximately the same number of the pattern occurrences per one session (more than 92 percent of sessions have one pattern) and the pattern lengths (preponderance of length three).

The most popular patterns in the whole semester for the direction from the grade page to the exam terms page are:

- 46,819 occurrences (52.03%)
  1. /auth/student/moje_znamky.pl
  2. /auth/index.pl
  3. /auth/student/prihl_na_zkousky.pl

- 21,307 occurrences (23.68%)
  1. /auth/student/moje_znamky.pl
  2. /auth/student/index.pl
  3. /auth/student/prihl_na_zkousky.pl

- 6,099 occurrences (6.78%)
  1. /auth/student/moje_znamky.pl
  2. /auth/student/prihl_na_zkousky.pl

- 1,284 occurrences (1.43%)
  1. /auth/student/moje_znamky.pl
  2. /auth/index.pl
  3. /auth/student/zapis.pl
  4. /auth/index.pl
  5. /auth/student/prihl_na_zkousky.pl

Script `/auth/ucitel/znamky.pl`

This is the application where teachers enter grades. They can variously filter the list of students before they put their grades into the system.

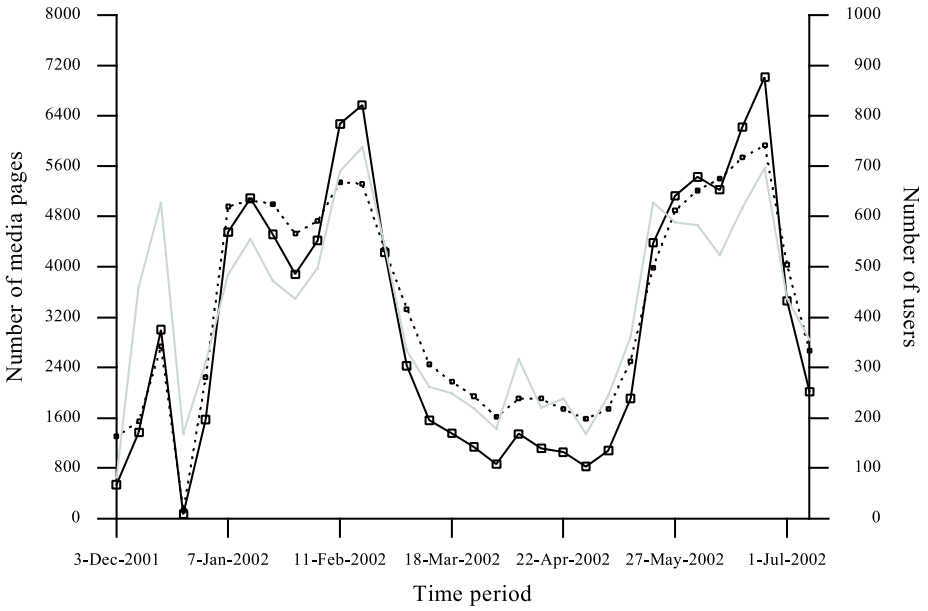

Figure 5.5: The `/auth/ucitel/znamky.pl` statistics

The graph shows the incremental rising of this application usage until the last week of the exam periods. At the maximums is the pages/users ratio almost ten pages per teacher. Furthermore, two interesting usage decreases can be seen at the end of January and in the middle of June.

The departure pages were again mostly the index pages – the `/auth/index.pl` with approximately one half of occurrences and the `/auth/ucitel/index.pl` with approximately one third of occurrences. Again, some users had this page in the bookmarks because in about 13 percent of occurrences this page was the first in a server session. Almost one half of sessions ended on this page whereas the others continued with one of the above index pages (about 40 percent) or with the `/auth/ucitel/zkusebni_terminy.pl` page (script for the exam terms administration). The detailed listings of the preceding and following pages are in Tables 5.20 and A.4.

Script name	Semes.	Regis.	Teach.	Exam
/auth/index.pl	49.17%	47.12%	54.47%	48.71%
/auth/ucitel/index.pl	32.25%	34.10%	30.86%	30.99%
none	13.54%	13.73%	10.30%	14.88%
/auth/ucitel/znamky_navod.pl	1.50%	1.71%	1.58%	1.27%
/auth/ucitel/zkusebni_terminy.pl	0.61%	0.56%	0.54%	0.71%
/auth/lide/foto.pl	0.54%	-	-	1.19%
/auth/ucitel/blok_edit.pl	-	0.54%	-	-
/auth/ucitel/znamky.pl	51,639	25,953	13,815	18,213
none	47.74%	46.01%	48.85%	49.51%
/auth/index.pl	23.89%	23.75%	26.20%	22.70%
/auth/ucitel/index.pl	17.16%	18.60%	15.12%	16.33%
/auth/ucitel/zkusebni_terminy.pl	4.00%	4.18%	2.78%	4.39%
/auth/ucitel/znamky_navod.pl	1.82%	2.08%	1.87%	1.54%
/auth/system/kill.pl	1.31%	1.40%	1.28%	1.19%
/auth/mail/mail.pl	0.96%	0.96%	1.09%	0.88%
/auth/ucitel/seznam_foto.pl	-	-	-	1.12%

Table 5.20: The `/auth/ucitel/znamky.pl` whence-where pages

The path analysis was performed on the last mentioned script. Approximately one sixth of sessions with the `/auth/ucitel/zkusebni_terminy.pl` script contain also the `/auth/ucitel/znamky.pl` script. Most sessions have only one pattern occurrence (more than 80 percent), seldom they have two (about 13 percent). The extreme is 23 occurrences which one user accomplished in the exam period. The patterns are usually three or four pages length.

	Semester		Registration		Teaching		Exam	
Number of all sessions	1,291,386		626,960		539,299		294,915	
/auth/ucitel/zkusebni_terminy.pl	17,305	1.34%	8,699	1.39%	4,324	0.80%	5,603	1.90%
/auth/ucitel/znamky.pl	36,068	2.79%	18,065	2.88%	9,845	1.83%	12,641	4.29%
Number of sessions with pattern	2,817	0.22%	1,522	0.24%	551	0.10%	1,122	0.38%
Number of all patterns	4,790		2,429		829		1,901	
1 occurrence	2,285	81.11%	1,255	82.46%	458	83.12%	904	80.57%
2 occurrences	377	13.38%	191	12.55%	73	13.25%	149	13.28%
Pattern length 2	317	6.62%	145	5.97%	74	8.93%	129	6.79%
Pattern length 3	1,965	41.02%	1,000	41.17%	362	43.67%	739	38.87%
Pattern length 4	1,752	36.58%	882	36.31%	274	33.05%	753	39.61%
Pattern length 5	355	7.41%	189	7.78%	57	6.88%	131	6.89%
Pattern length 6	166	3.47%	82	3.38%	31	3.74%	62	3.26%

Table 5.21: the `/auth/ucitel/zkusebni_terminy.pl` → `/auth/ucitel/znamky.pl` patterns

The most popular patterns in the whole semester are:

- 1,275 occurrences (26.62%)
  1. /auth/ucitel/zkusebni_terminy.pl
  2. /auth/ucitel/seznam.pl
  3. /auth/ucitel/index.pl
  4. /auth/ucitel/znamky.pl

- 1,051 occurrences (21.94%)
  1. /auth/ucitel/zkusebni_terminy.pl
  2. /auth/index.pl
  3. /auth/ucitel/znamky.pl

- 868 occurrences (18.12%)
  1. /auth/ucitel/zkusebni_terminy.pl
  2. /auth/ucitel/index.pl
  3. /auth/ucitel/znamky.pl

- 317 occurrences (6.62%)
  1. /auth/ucitel/zkusebni_terminy.pl
  2. /auth/ucitel/znamky.pl

- 227 occurrences (4.74%)
  1. /auth/ucitel/zkusebni_terminy.pl
  2. /auth/ucitel/seznam.pl
  3. /auth/index.pl
  4. /auth/ucitel/znamky.pl

Chapter 6

Conclusion

Our work concerns the Web Usage Mining. We put it in the context of Web Mining and describe all parts of the mining process. It begins with the data sources identification when we can choose from three data types – server-level, client-level, or proxy-level collections. Next, the process continues with the preprocessing stage where the most important part takes place – the usage preprocessing. It creates mineable objects (server sessions or episodes) for pattern analysis and discovery. The pattern analysis utilizes many techniques based on statistics, machine learning, pattern recognition and the like. Although there are many commercial analysis tools, we can generally use any tool which processes and filters the discovered patterns. At the end of the first part we describe the application areas of the Web Usage Mining.

The second part of this thesis concerns our practical tests of the Web Usage Mining methods on the data from the Information System of Masaryk University. First, we dealt with a huge amount of data. The IS is an extensive system and it is intensively used. The data for mining (with indexes) occupies around 10 GB in the database.

First, we present statistics for one semester. The statistics confirm a high usage of the IS during the registration and exam periods. This is not surprising because the biggest group of IS users are students who use the course registration and exam terms agenda during these periods. Next, it shows interesting results about the long sessions – the majority of them probably violate the IS photo policy. Finally, it shows that users usually access around ten pages per session – they check some information and leave the IS.

Then we present a detailed path analysis of four applications. We present how these applications are used during the semester, show the usual departure and following pages of these applications, and the most popular patterns

between them and one more page. The developed framework can be applied to any application of the IS. The results of this part confirmed our beliefs of the IS usage. Users travel directly to their targets. Many users also utilize bookmarks (or similar tool) to achieve the target pages in one click.

Possible extensions of our work include the analyses of the non-authenticated users or the script parameters. Also, an application offering more user friendly querying of the current database could be developed.

Bibliography

[1] Informační systém Masarykovy univerzity v roce 2002. `http://is.muni.cz/clanky/vyrocka2002.pl`.

[2] Information system MU. `http://is.muni.cz`.

[3] NetZero. `http://www.netzero.net`.

[4] Spedia. `http://www.spedia.net`.

[5] SurfAid. `http://surfaid.dfw.ibm.com`.

[6] T. Berners-Lee, R. Fielding, U.C. Irvine, and L. Masinter. Uniform resource identifiers (URI): Generic syntax. `http://www.rfc-editor.org/rfc/rfc2396.txt`, 1998.

[7] Alex G. Buchner and Maurice D. Mulvenna. Discovering Internet marketing intelligence through online analytical web usage mining. *SIGMOD Record*, 27(4):54–61, 1998.

[8] L. Catledge and J. Pitkow. Characterizing browsing behaviors on the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[9] M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Sixteenth International Conference on Distributed Computing Systems*, pages 385–392, 1996.

[10] R. Cooley. Web usage mining: Discovery and application of interesting patterns from web data. URL: `http://citeseer.nj.nec.com/article/cooley00web.html`, 2000.

[11] R. Cooley, J. Srivastava, and B. Mobasher. Web mining: Information and pattern discovery on the world wide web. URL: `citeseer.nj.nec.com/cooley97web.html`, 1997.

[12] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[13] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1995.

[14] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[15] R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical report CS 99-04, Department of Computer Science, University of Regina, October 1999.

[16] Thorsten Joachims, Dayne Freitag, and Tom M. Mitchell. Web watcher: A tour guide for the world wide web. In *IJCAI (1)*, pages 770–777, 1997.

[17] B. Lavoie and H. F. Nielsen. Web characterization terminology and definitions sheet. `http://www.w3.org/1999/05/WCA-terms/`, 1999.

[18] A. Luotonen. The Common Logfile Format. `http://www.w3.org/Daemon/User/Config/Logging.html`, 1995.

[19] Sanjay Kumar Madria, Sourav Sourav Bhowmick, Wee Keong Ng, and Ee-Peng Lim. Research issues in web data mining. In *Data Warehousing and Knowledge Discovery*, pages 303–312, 1999.

[20] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *IEEE Knowledge and Data Engineering Workshop (KDEX'99)*, 1999.

[21] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*. ACM Press, 1996.

[22] Ellen Spertus. Parasite: Mining structural information on the web. *Computer Networks and ISDN Systems: The International Journal of Computer and Telecomunication Networking*, 29:1205–1215, 1997.

[23] Myra Spiliopoulou and Lukas C. Faulstich. WUM: a Web Utilization Miner. In *Workshop on the Web and Data Bases (WebDB98)*, pages 109–115, 1998.

[24] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

Appendix A

Five-distance whence-where pages

Script name	Semes.	Regis.	Teach.	Exam
none	44.74%	43.81%	45.61%	45.43%
/auth/index.pl	26.66%	26.79%	26.35%	26.74%
/auth/student/index.pl	5.51%	5.68%	5.25%	5.54%
/auth/student/prihl_na_zkousky.pl	4.32%	4.38%	4.76%	3.55%
/auth/mail/index.pl	3.84%	3.67%	4.25%	3.64%
/auth/student/moje_znamky.pl	3.75%	3.94%	2.36%	5.31%
/auth/mail/mail.pl	3.70%	3.60%	3.98%	3.49%
/auth/student/zapis.pl	1.03%	1.24%	0.83%	0.97%
/auth/nastenka/zprava.pl	0.65%	0.59%	0.89%	-
/auth/mail/mail_posli.pl	0.62%	0.60%	0.69%	0.59%
/auth/lide/index.pl	0.59%	0.61%	0.62%	0.50%
/auth/studium_spolec/index.pl	-	0.51%	-	-
/auth/student/prihl_na_zkousky.pl	2,662,375	1,300,265	891,465	664,695
none	58.74%	56.82%	60.20%	60.49%
/auth/index.pl	14.44%	14.87%	14.32%	13.63%
/auth/student/index.pl	4.50%	4.92%	3.97%	4.37%
/auth/student/prihl_na_zkousky.pl	4.32%	4.38%	4.76%	3.55%
/auth/student/moje_znamky.pl	3.43%	3.54%	2.47%	4.61%
/auth/mail/index.pl	1.91%	1.81%	2.15%	1.78%
/auth/mail/mail.pl	1.83%	1.79%	1.98%	1.64%
/auth/student/zapis.pl	1.53%	1.85%	1.11%	1.68%
/auth/lide/index.pl	1.16%	1.17%	1.20%	1.10%
/auth/lide/foto.pl	0.82%	0.81%	0.86%	0.79%
/auth/studium_spolec/index.pl	0.69%	0.83%	0.51%	0.70%
/auth/seminare/student.pl	0.51%	0.60%	-	0.59%
/auth/mail/mail_posli.pl	0.50%	-	0.56%	0.51%
/auth/dok/index.pl	-	-	0.60%	-

Table A.1: The `/auth/student/prihl_na_zkousky.pl` five-distance whence-where pages

Script name	Semes.	Regis.	Teach.	Exam
none	38.17%	38.58%	38.15%	36.78%
/auth/index.pl	28.07%	27.16%	28.44%	28.69%
/auth/mail/index.pl	4.56%	4.03%	5.35%	3.78%
/auth/mail/mail.pl	4.32%	3.85%	4.98%	3.61%
/auth/dok/index.pl	4.25%	4.63%	4.43%	3.54%
/auth/student/prihl_na_zkousky.pl	3.39%	2.89%	3.07%	3.91%
/auth/student/moje_znamky.pl	1.90%	2.16%	1.21%	3.60%
/auth/student/index.pl	1.74%	1.63%	1.49%	2.72%
/auth/nastenka/zprava.pl	1.25%	0.99%	1.65%	0.61%
/auth/student/zapis.pl	1.23%	1.56%	1.01%	1.47%
/auth/lide/index.pl	1.08%	1.06%	1.13%	1.00%
/auth/lide/foto.pl	0.84%	0.84%	0.88%	0.78%
/auth/mail/mail_posli.pl	0.77%	0.72%	0.91%	0.60%
/auth/seminare/student.pl	0.68%	0.71%	-	1.23%
/auth/dok/dok_novy.pl	0.59%	0.80%	0.58%	-
/auth/predmety/predmet.pl	0.58%	0.79%	-	0.58%
/auth/studium_spolec/index.pl	0.52%	0.59%	-	0.70%
/auth/dok/index.pl	302,290	131,010	174,020	44,735
none	61.32%	58.84%	62.79%	61.67%
/auth/index.pl	14.46%	15.08%	14.14%	14.04%
/auth/dok/index.pl	4.25%	4.63%	4.43%	3.54%
/auth/mail/index.pl	1.98%	1.80%	2.17%	1.90%
/auth/mail/mail.pl	1.91%	1.79%	2.04%	1.77%
/auth/student/prihl_na_zkousky.pl	1.70%	1.59%	1.49%	1.78%
/auth/lide/index.pl	1.29%	1.35%	1.31%	1.16%
/auth/student/index.pl	1.18%	1.18%	0.99%	1.73%
/auth/student/moje_znamky.pl	1.13%	1.32%	0.79%	1.79%
/auth/lide/foto.pl	0.93%	0.97%	0.93%	0.89%
/auth/student/zapis.pl	0.85%	1.10%	0.68%	0.93%
/auth/dok/dok_novy.pl	0.65%	0.87%	0.62%	-
/auth/mail/mail_posli.pl	0.59%	0.55%	0.68%	-
/auth/studium_spolec/index.pl	0.52%	0.57%	-	0.76%
/auth/seminare/student.pl	-	-	-	0.70%
/auth/predmety/predmet.pl	-	0.62%	-	-
/auth/predmety/index.pl	-	0.58%	-	-
/auth/nastenka/zprava.pl	-	-	0.56%	-

Table A.2: The `/auth/dok/index.pl` five-distance whence-where pages

Script name	Semes.	Regis.	Teach.	Exam
none	19.33%	19.75%	19.67%	18.91%
/auth/predmety/index.pl	16.01%	13.85%	15.97%	18.90%
/auth/index.pl	12.20%	13.68%	11.32%	11.93%
/auth/predmety/katalog_plneni.pl	11.63%	9.45%	12.81%	12.06%
/auth/predmety/uplny_vypis.pl	8.10%	8.66%	8.20%	6.61%
/auth/predmety/atributy_plneni.pl	5.82%	2.06%	8.14%	7.35%
/auth/predmety/sylaby_plneni.pl	4.44%	5.23%	3.78%	3.61%
/auth/predmety/predmet.pl	2.40%	3.48%	1.93%	1.55%
/auth/predmety/katalog_vypis.pl	2.34%	2.43%	2.95%	1.60%
/auth/ucitel/index.pl	1.36%	1.97%	1.08%	1.28%
/auth/lide/foto.pl	1.35%	1.44%	1.39%	1.39%
/auth/studium/zapis.pl	1.20%	1.02%	0.99%	2.02%
/auth/predmety/katalog_plneni_hromadne.pl	0.80%	0.54%	0.87%	0.96%
/auth/ucitel/znamky.pl	0.76%	0.83%	0.66%	1.05%
/auth/studium_spolec/index.pl	0.71%	0.94%	-	0.82%
/auth/lide/index.pl	0.71%	1.00%	0.61%	-
/auth/predmety/sablony_plneni.pl	0.70%	0.69%	-	1.10%
/auth/predmety/katalog.pl	0.68%	0.78%	0.59%	0.77%
/auth/student/zapis.pl	0.61%	1.06%	-	-
/auth/ucitel/seznam.pl	-	0.53%	-	-
/auth/prezentator/index.pl	-	-	0.65%	-
/auth/predmety/sablony.pl	-	0.50%	-	-
/auth/predmety/katalog_plneni.pl	74,895	30,670	34,205	17,240
none	27.65%	30.17%	27.05%	26.32%
/auth/predmety/katalog_plneni.pl	11.63%	9.45%	12.81%	12.06%
/auth/predmety/index.pl	11.30%	8.31%	11.86%	14.19%
/auth/index.pl	7.90%	9.44%	6.79%	7.85%
/auth/predmety/uplny_vypis.pl	7.64%	7.92%	7.91%	6.31%
/auth/predmety/atributy_plneni.pl	7.21%	2.91%	9.88%	8.87%
/auth/predmety/sylaby_plneni.pl	5.74%	7.07%	4.89%	4.36%
/auth/predmety/katalog_vypis.pl	1.92%	1.91%	2.45%	1.22%
/auth/predmety/predmet.pl	1.77%	2.43%	1.52%	1.18%
/auth/studium/zapis.pl	1.36%	1.16%	1.20%	2.20%
/auth/ucitel/index.pl	1.15%	1.65%	0.92%	1.02%
/auth/lide/foto.pl	1.13%	1.12%	1.24%	1.18%
/auth/studium_spolec/index.pl	0.91%	1.22%	-	1.05%
/auth/predmety/katalog_plneni_hromadne.pl	0.84%	0.60%	0.90%	1.07%
/auth/lide/index.pl	0.79%	1.00%	0.72%	0.60%
/auth/predmety/sablony_plneni.pl	0.75%	0.78%	-	1.14%
/auth/ucitel/znamky.pl	0.64%	0.72%	0.58%	0.82%
/auth/predmety/katalog.pl	0.60%	0.68%	0.54%	0.71%
/auth/ucitel/seznam.pl	-	0.54%	-	-
/auth/studium/index.pl	-	0.52%	-	-
/auth/student/zapis.pl	-	0.66%	-	-
/auth/prezentator/index.pl	-	-	0.84%	-

Table A.3:
The `/auth/predmety/katalog_plneni.pl` five-distance whence-where pages

Script name	Semes.	Regis.	Teach.	Exam
none	49.89%	49.01%	49.98%	50.97%
/auth/index.pl	19.01%	18.77%	20.80%	18.37%
/auth/ucitel/index.pl	11.06%	11.81%	10.62%	10.47%
/auth/ucitel/znamky.pl	6.49%	6.55%	6.40%	6.51%
/auth/ucitel/zkusebni_terminy.pl	2.40%	2.41%	1.55%	2.71%
/auth/ucitel/seznam.pl	2.25%	2.40%	2.01%	2.18%
/auth/lide/foto.pl	1.08%	0.93%	0.80%	1.40%
/auth/mail/index.pl	0.76%	0.74%	0.78%	0.78%
/auth/mail/mail.pl	0.55%	0.53%	0.58%	0.54%
/auth/ucitel/znamky_navod.pl	-	0.53%	-	-
/auth/ucitel/seznam_foto.pl	-	-	-	0.71%
/auth/ucitel/blok_edit.pl	-	0.52%	-	-
/auth/nastenka/zprava.pl	-	-	0.56%	-
/auth/ucitel/znamky.pl	258,195	129,765	69,075	91,065
none	61.52%	60.04%	63.17%	62.70%
/auth/index.pl	9.15%	9.32%	9.77%	8.50%
/auth/ucitel/index.pl	7.64%	8.36%	6.78%	7.19%
/auth/ucitel/znamky.pl	6.49%	6.55%	6.40%	6.51%
/auth/ucitel/zkusebni_terminy.pl	2.59%	2.68%	1.73%	2.81%
/auth/ucitel/seznam.pl	1.82%	2.06%	1.53%	1.68%
/auth/lide/foto.pl	1.17%	1.04%	1.02%	1.41%
/auth/lide/index.pl	0.70%	0.74%	0.77%	0.64%
/auth/mail/index.pl	0.62%	0.62%	0.68%	0.59%
/auth/ucitel/seznam_foto.pl	0.59%	-	-	0.90%
/auth/mail/mail.pl	0.53%	0.53%	0.60%	-
/auth/ucitel/znamky_navod.pl	0.50%	0.58%	0.52%	-
/auth/ucitel/statistika_znamek.pl	-	-	-	0.57%
/auth/ucitel/blok_edit.pl	-	0.50%	-	-
/auth/system/kill.pl	-	0.52%	-	-
/auth/predmety/index.pl	-	-	0.53%	-

Table A.4: The `/auth/ucitel/znamky.pl` five-distance whence-where pages