Applying User Profile Ontology for Mining Web Site Adaptation Recommendations

Tarmo Robal, Ahto Kalja

Dept. of Computer Engineering, Tallinn University of Technology
Raja 15, 12618 Tallinn, Estonia
tarmo@pld.ttu.ee, ahto@cs.ioc.ee

Abstract. The Internet consists of web sites that employ different kinds of structures as the backbone of their build-up. However, users are browsing the web according to its content, regardless of the structure. In this paper we discuss the possibilities of applying ontologies in exploring the web sites' structures and usage for producing viewing recommendations for the visitors. A special log system for capturing access data is introduced as well as techniques applied for data mining. Ontology of user profiles is constructed by exploiting the user locality model.

Keywords. Web mining, web log, web ontology, recommender systems.

1 Introduction

The World Wide Web as an endless source of information has become customary to our everyday lives. The information is mainly represented in the form of web pages that users may browse; and the way they do it is heavily dependent on the information they are searching for. It has been proven that factors, such as visual experience and site attractiveness, logicality of navigation organization (especially on large sites), placement of objects, colour schema and page loading time [1,2,3,4] also affect web surfing. Furthermore, the preferences of users are changing through time. Thus, there exists a constant need for improving and adapting applications and services according to the end-users' needs. For that we need to collect access data and mechanisms for data mining and processing. The rapid expansion of the Internet has provided the possibility to explore users' navigational patterns and interaction with web-based systems. This allows computing of recommendations for users, either by simple pages, or constructing new web site topologies for user groups.

In this article we discuss generation of recommendations for website adaptation based on user locality model and website ontology. A locality model, user profiles ontology mining and recommendations engine with reference to the practical merit to be applied on web sites are proposed. For empirical studies, a real website access data, collected from the usage of the website of the Department of Computer Engineering (DCE) at Tallinn University of Technology (TUT), has been used. The paper is organized as follows. In section 2, we discuss web mining, adaptive web sites

Y. Ioannidis, B. Novikov and B. Rachev (Eds.): Local Proceedings of ADBIS 2007, pp. 126-135

and recommender systems in general with emphasis on the aspects important in this paper. Also, an overview of data collection methodologies with respect to the log system used to capture and analyse web usage data, as well as the procedures implemented for data preparation and profiles extraction are discussed. In Section 3, we consider recommendation generation based on active user's locality discovery and mapping of user profiles ontology to web ontology. Section 4 draws some conclusions.

2 Web Mining and Adaptive Web Sites

Collecting user interaction data during the web sessions and analysing it in further, enables to develop assistance systems to help users during their web browsing, either directly (e.g. providing related links) or indirectly, i.e., constantly improving and adapting the website to the users' needs. However, usually such analysing systems, as also noted in [5], suffer from inability to address the semantic aspects of the web. The use of ontologies is an emerging trend in web mining, as it is one of the best ways for representing the structure of information regarding to its semantics. Herewith, we apply ontologies to give meaning to the results of our data mining process.

2.1 Web Mining

The web mining process is usually divided into three categories: (1) web content mining, (2) structure mining and (3) usage mining [6]. Some authors also highlight a fourth category called user profile mining [7]. Srivastava et al. [8] have studied research projects and products in the field and have outlined the major application areas as follows: web personalization, system improvement (performance and other service quality attributes to guarantee user satisfaction), site modification (attractiveness of a website in terms of content and structure), business intelligence (applying web access data to discover marketing trends for e-commerce applications), and usage characterization (user interaction with the browser interface and navigational strategies for browsing a particular web-site). The domain is quite wide-scoped. Hereby, we will concentrate on usage and structure mining for constructing recommendations based on predefined user profiles and web ontology.

2.2 Adaptive Websites and Ontologies

Web structure mining is engaged to the structure of the web site and therefore it is the basis for adaptive websites. Adaptive web sites are defined as those, which are strategically or tactically modified according to their usage. The web site topology construction and restructuring is useful to build upon the site's ontology as then and only then the concepts of that particular website can be involved. The adaptation process can be either: tactical, where adaptations are triggered in real time and do not need the approval of a webmaster, as they do not interfere the structure of the site whilst completing it with additional information; or strategic, where adaptations can

have a serious impact on the sites structure; therefore they need to be done offline and with the approval of webmaster.

The semantic aspects of the adaptations of website should also be taken into account. While in [9] the adaptations are divided into two categories: short-term (tactical) and long-term (strategic) changes, and an algorithm for self-adaptive website is discussed, the issue of semantics behind the individual pages of which the site consists, is not considered. Nevertheless, the framework provides good basis for semantic web adaptation. As proposed in [5], website ontology is strongly related to the site's topology, as the ontology is comprised of the thematic categories covered by the pages of the web site; where each page is an instance of one or more concepts. These concepts can be arranged in a hierarchy with "is-a" relationships. Moreover, if concepts are defined in a way that each instance exists in the form of a web page, there will also exist a relationship instance in the form of web page pair [6].

For web site adaptation, we need the original web ontology constructed by the webmaster and an activity log of users' operations on the site, from which the modalities of the site access can be extracted. As proposed by Lim and Sun [10], having web pages as concept instances and applying ontology-based structure mining, it is possible to derive linkage patterns among concepts from web pages for the site's design improvement. Hereby, the design is noted as the site's structure. Such an improvement can also be a refined and recommended topology constructed for enhanced viewing experience.

2.3 Recommender Systems

The ability to track down users' actions has made it possible to develop systems which provide users with dynamically discovered recommendations and thus personalize the users' web experience. The aim of any recommender system is to assist users to find needed information in the easiest way via helping them to discover pages or page-sets they otherwise might not find during their site visit. Moreover, additional related information, not contained on that particular web site but which users might also find relevant, can be provided. For employing such systems, usage data has to be collected and processed. As a result, tactical adaptation of web sites can be implemented for enhanced and improved user experience.

In general, recommender systems are based on site's overall usage information and therefore transparent for the end-users or being personalized with the aim to adjust the web cognition for a particular user. The latter though needs users to log in to a system or other means of explicit user identification. Recommendation systems are being already successfully used in personal web-based agents such as Letizia, Syskill&Webert, Personal Webwatcher, and OntoSeek. They have also been implemented for personalised e-learning, as discussed in [11].

2.4 Data Collection Methods

Data can be gathered using either explicit or implicit methods. The first assumes users to actively participate in data gathering. For example, we might ask users to rate

pages on a sliding scale or give pages credits like in [12] and [13], where an evaluation form at the end of web pages is presented and users are asked to rank the pages with the possibility to add comments. Another option is to have surveys for web sites, such as [14]. These surveys though contain far more questions, which usually are not page-specific. Applying explicit data collection methods is a useful technique and provides the users' opinions and ideas; however users are usually not willing to actively participate in such evaluations, moreover fill in forms.

Implicit techniques, being transparent to the end-users, help to overcome these disadvantages. Automated data gathering has enabled monitoring of accessed pages, navigational paths; discovery of usage patterns and user profiles. The techniques usually involve data collection from server, proxy or client level using either web server logs, web browsers modified for data capturing, or special log systems [8].

Web server logs chronicle all the operations and do not produce log data for particular analyses. Most probably we are not interested in every single object accessed on a web server (e.g. dots and lines as elements of graphic design), which is also the case of this study. Moreover, it has been proven in [15,18] that HTTP traffic logs appear to be flawed and there are some major difficulties due to data incompleteness [8,16]. One of the many problems with web server log files is that they do not allow to identify visitor sessions [17].

As shown, standard web server logs occurred to be not suit-able, as they did not contain all the needed information, for instance clustering of user actions into sessions and tracking recurrent visits. After analysing different possibilities of usage data collection and taking into account the aforementioned drawbacks, it was clear that a different kind of approach was in demand to meet the needs of this study. As a result, a special log system was used (Section 2.5).

2.5 The Log System

The log system [18] was developed in 2003 and based on a preliminary log system introduced in 2002, which allowed to store only some basic properties of actions users performed and was only aimed on general usage statistics. The major improvement towards the new log system was the ability to capture distinct and recurring user sessions, which is also the basis of users' profiles construction.

The web access log contains raw access data, which needs to be cleaned from noise and filtered before it can be used for user profiles extraction. Therefore, crawlers/robots detection rules were added to the log system, based on [19,20,21] and our own experience, as blocking of search robots incorporates the risk of loosing potential visitors. Moreover, most of the robots gather and index the content of the site for free. The log system is initialised every time a page request is made (Fig.1). The approach is totally transparent and thus requires no active participation from the users. Both, the log system and analyser system are based on MySQL DBMS [22], the first on MyISAM storage engine and the latter on InnoDB.

Currently the log is about 189.5 MB in raw data size, consisting of more than 1 022 984 records (269 782 sessions) captured from the DCE website over the time period of 5 years. DCE website log was chosen for data analyses demonstration as it has been used for the longest period of time.

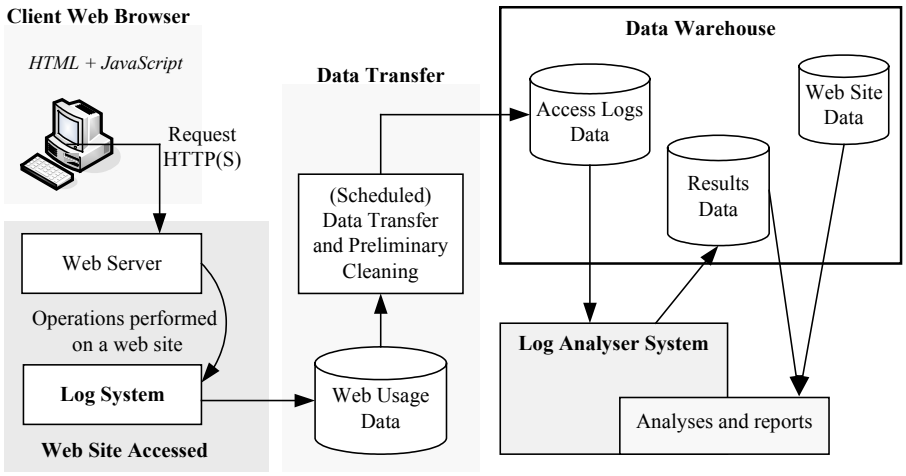

Fig. 1. Data collection and preprocessing using the log system and analyser.

From the data captured by the log system [18], we used the following: (1) requested page, (2) client identifier (session based ID), (3) operations performed during a session.

2.6 User Profiles Extraction Based on Locality Model

The raw web access data stored in the log was cleaned from noise and filtered before it was used for user profiles extraction. During the process, user navigational paths were composed of unique identifiers for web pages from session data (Fig. 2). The analyser system derived from the log 269 782 navigational paths of which a total of 87 953 remained to be processed in further after filtering.

The locality model applied here is based on the belief that if a large number of users frequently access a set of pages, then these pages must be related. The locality L is defined as the users' nearest sequential activity history within the site. Thereby, during the web session users are moving from one locality to another, which can be represented by the w latest operations (accesses to pages) they made.

Let us now discover the appropriate size of the sliding window $w \in W$. For that, localities with various window sizes were constructed based on the discovered navigational paths. The results of the empirical study were evaluated by the following properties: (1) cover percentage for the number of combinations computed from the paths, (2) average frequency of finding these combinations in paths, (3) average number of possible localities in path, and (4) the availability of next item for each locality.

A user session is defined as a sequence of accessed pages $s=<p_i, p_{i+1},\ldots p_n>$, where $p_i \in P$ and P is the set of all pages. Thus we are searching for localities L of size w, such as $L=p_j, p_{j+1}, \ldots p_m$, where $p_j \neq p_{j+1} \neq \ldots \neq p_m$, p_j is a visited item (page ID). For each $s_t \in S$ we apply a function $L=CalculateLocality(s_t,w)$. The attribute "next item" is found as p_{j+1}, if available.

Table 1. Results of empirical window size *w* study.

Properties observed	Studied window size w			
	2	**3**	**4**	**5**
(1) Combination coverage [%]	31.2	35.5	20.7	12.6
(2) Combination frequency	1.1	1.0	1.0	1.0
(3) No of localities in path	6.3	6.6	6.5	5.9
(4) Availability of next item [%]	76.6	77.4	74.1	76.3

After conducting the experiments with W={2,3,4,5} (Table 1) it became clear that the size of the sliding window might be in correlation with the absolute menu depth of the web site. Interestingly, the menu structure of DCE has three levels, thus w=3. As can be seen from the table, locality with window size w=3 performed the best, according to the properties observed and considering the fact that w=2 would be too short for describing recent actions in sense of user profiling.

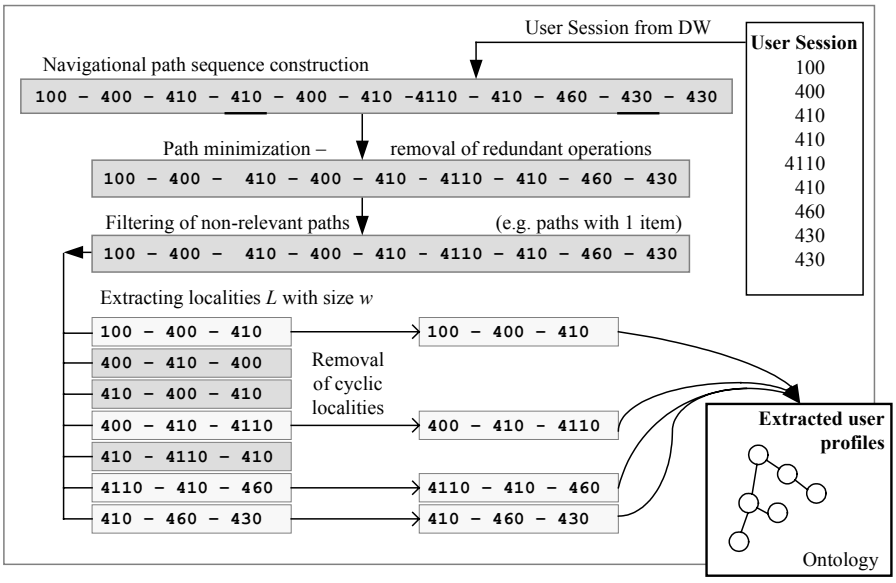

Fig. 2. The process of user profiles extraction from navigational paths.

Having determined the suitable value for *W*, the localities *L* were further filtered to remove such $L= p_1p_2..p_w$, where $p_1=p_w$., as they represented cyclic localities (during the locality users end up at the same point they started), which is an interest of another study. For instance, we removed the locality "400-410-400".

Obviously all the remaining localities do not represent general trends and we have to take into account their frequency in *L*. Infrequent localities are probably the result of random operations. Thus, only the top frequent localities represent the user profiles. In further *L* is referred as *extracted user profiles*, which are used in ontology to compute recommendations.

3 Recommendations Based on Online User Profile Discovery

The amount of information available over the Internet is enormous - even the quantum of information on an average web site is fairly large for users to maintain easy navigation. This problem can be revealed by providing recommendations based on user profiles mined from web logs. The rationale behind the approach of recommender systems is that users implicitly use a concept model based on their own knowledge of the domain or topic searched, even though mostly they do not know how to represent it [7]. However, having unobtrusively monitored users' actions and collected them in the log, it is possible to apply detection algorithms to produce the concept models of users, i.e., extract user profiles.

By analysing the actions of the current user online and comparing it to the user profiles discovered from the web log, it is possible to classify the user as an individual into one of the conceptual user groups and recommend him/her new pages that correlate with that individual (tactical adaptation), highlight items, so that they could be easily located or even propose a new topology - strategic adaptation (Fig. 3.).

The items in extracted user profiles do not contain any semantic information about the content represented on the web-site. Therefore, the extracted user profiles need to be mapped to the concepts of web ontology. For instance extracted user profile "400-410-4110" withholds the concepts "Students", "List of Subjects", and "Subject Content". For that reason, profiles ontology is constructed. At this point, it was composed manually, however in the future it is planned to implement an automated generation of the OWL-file. In the user profile ontology, several user concepts were defined in order to give meaning to the mining results and propose recommendations. For instance, concepts "Researcher" and "Student". The user profiles ontology is mapped to the web ontology (Fig. 3).

While building the web site ontology, we have to posit ourselves on the users' side – they behave according to the view they have on a display, and make requests for a web page not a concept. Then and only then we are able to use the concepts of web ontology to characterize the actions of users and map semantics to profiles for producing recommendations that are semantically related to current user's session based on the locality model.

The task of the recommendation engine (RE) is to determine the type of the user online and compute recommendations based on the recent actions of that user. The decision is based on the knowledge attained from the ontologies and page ranking. No user is attached to a particular outcome, as users are free to move from one locality to another. However, with carefully designed user profile ontology, the recommendation covers the majority of the user session. The pages are ranked using an inverse time weighting algorithm (1) [23]:

$$\text{Rank} = p\sum_{i=1}^{n} \frac{\text{Interest value (i)}}{\text{Age (i)}}. \tag{1}$$

In the formula, *Age(i)* represents number of days into the past, *Interest value (i)* number of hits for a page during *Age(i)*, *p* is a probability value between 0 and 1 and can be predefined. Thus, only the nearest past will play a crucial role in ranking.

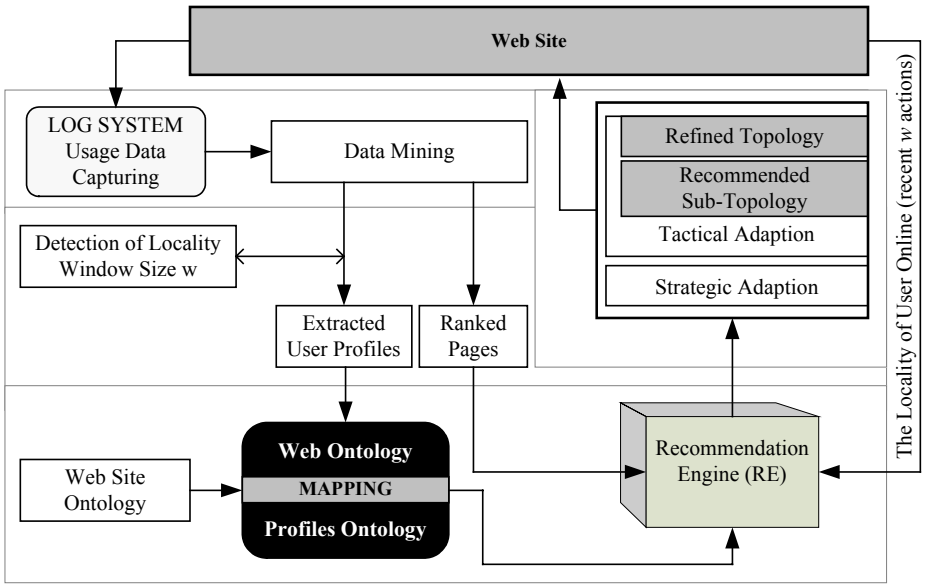

Fig. 3. Recommendations generation based on extracted user profiles.

In terms of the RE, let us consider profiles ontology with concepts "*Department - Students - ListOfSubjects*", "*Research - Publications - Partners*", "*Students - ListOfSubjects - SubjectContent*" in it. During the recommendation generation process concepts as "*Department - Students - ListOfSubjects*", "*Students - ListOfSubjects - SubjectContent*" are reasoned to belong to a user profile defined as "*Student*". Thus, we know that users profiled as students are interested in these concepts of the web ontology, which belong to those extracted profiles. The discovered knowledge can now be applied for tactical or strategic adaptation.

3.1 Tactical Recommendations

The tactical adaptation involves a refined topology by adding items to an existing web topology in a way that they would be more easily accessed by users. In its simplest case, the recommended items in an existing web site topology are raised during the user's online session. For instance, if the user is classified to be a student, then the set of items to be added into topology or highlighted in the existing one would consist of items belonging to those concepts identified by RE and with ranking applied.

Setting up a new web site topology as a result of tactical adaptation, would be an option as well. However, we must not forget that there is a risk to confuse the end-users by changing the topology too often. Additionally, users might be provided an option to classify themselves into one of the users' concepts by clicking a special link. An option to return to a regular web site based on the full topology is a must. In the long run, reasoning about the recommendations made for tactical adaptations may propose strategic ones as well. For either case, the recommendation engine and ontologies are needed.

3.2 Strategic Recommendations

In terms of strategic adaptation, general site improvement is considered. The result of strategic recommendation is a new web site topology. As strategic adaptations need the approval of webmaster, they should not be applied online and should not be put into effect too often, as it may distress the regular visitors to find a new layout of the site every time they return. Nevertheless, web sites should adjust over time to the preferences of their users and therefore strategic changes are in need as well.

Analysing the log, sets of pages that are frequently used together, are to be recognized. Running a comparison on the existing web ontology and the results obtained from the actual use of the site, proposals for improvement can be made. The domain of interest covers sets of pages that are related in the ontology but not used together, and sets of pages that are used together but are not described in the web ontology. As can be expected, new relationships not existing in the current web ontology can be discovered. These newly discovered relationships may lead to the need to add links between sets of pages in order to improve content availability for the users. Moreover, over time the sets of pages discovered may impose new web ontology, thus a new structure to be developed.

4 Conclusions and Future Plans

The enormous amount of information available over the Internet has produced the need for recommender systems, which would help users to find the information they are seeking easily. Using various implicit data capturing methods, it is possible to track down users' actions and advise them to visit other related pages or assist in finding information according to their previous actions.

In this paper we have developed an approach for recommendations construction for web site improvement as tactical adaptations. The proposed methodology involves data mining from web access logs, construction of users' navigational paths and applying of locality model for user profiles extraction. The model is based on the recent actions performed by a user. These profiles are collected into ontology which together with web ontology is used to give meaning to the mined data. Applying the mapped ontology and the locality of the active session, the latter is classified into one or more predefined user concepts. As a result, recommendations for tactical adaptations can be generated by the recommendation engine. The latter can also be used for proposing strategic adaptations; going further, strategic changes can be a result of mined tactical recommendations.

Acknowledgements. We appreciate the support of Estonian Information Technology Foundation, Doctoral School in ICT of Measure 1.1 of the Estonian NDP and the Estonian Scientific Foundation grant no. 5766.

References

1. Bernard, M. L. (2001). User expectations for the location of web objects. Proceedings of CHI '01 Conference: Human Factors in Computing Systems, pp. 171-172. http://psychology.wichita.edu/hci/projects/CHI%20web%20objects.pdf [2007]
2. Geissler, G., et al. Web Home Page Complexity and Communication Effectiveness. Journal of the Association for Information Systems, Vol.2, Art. 2, 2001, pp. 1-48.
3. Bernard, M. L., & Chaparro, B. S. Searching within websites: A comparison of three types of sitemap menu structures. Proceedings of The Human Factors and Ergonomics Society 44th Annual Meeting in San Diego, 2000, pp. 441-444. (PDF format) http://psychology.wichita.edu/hci/projects/sitemap.pdf [2007]
4. Lee, A.T. Web usability: a review of the research, ACM SIGCHI Bulletin, Vol.31, Iss.1 1999, pp: 38 - 40
5. Mikroyannidis, A., Theodoulidis, B. Web usage Driven Adaptation of the Semantic Web, Proceedings of UserSWeb: Workshop on End User Aspects of the Semantic Web; Heraklion, Crete, May 29, 2005, pp. 137-147
6. Kosala, R., Blockeel, H., Web Mining Research: A Survey, ACM SIGKDD Explorations, Vol. 2, No. 1, 2000, pp. 1-15
7. Li, Y., Zhong, N. Mining Ontology for Automatically Acquiring Web User Information Needs, IEEE Trans. on Knowledge and Data Engineering, Vol. 18, No. 4, 2006, pp. 554-568
8. Srivastava, J. et.al., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, Vol. 1(2): 2000. pp. 12-23,
9. Coenen, F., Swinnen, G., Vanhoof, K., Wets, G. A Framework for Self Adaptive Websites: Tactical Versus Strategic Changes. In Proc. of WEBKDD'2000 Web Mining for E-Commerce – Challenges and Opportunities, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining
10. Lim, E-P., Sun, A. Web Mining – the Ontology Approach, International Advanced Digital Library Conference (IADLC'2005), Nagoya University, Nagoya, Japan, http://iadlc.nul.nagoya-u.ac.jp/archives/IADLC2005/Ee-Peng.pdf [2006]
11. Markellou, P., Mousouroulli, I., Spiros, S., Tsakadilis, A. Using Semantic Web Mining technologies for Personalized e-Learning Experiences, In Proc. of The IASTED International Conference on Web-based Education (WBE 2005) Eds: V. Uskov. 700 p.
12. IBM website, http://www.ibm.com/ [2007]
13. Hewlett-Packard website http://www.hp.com/ [2007]
14. Help Us Make NVIDIA.com Better http://www.nvidia.com/object/IO_27690.html [2007]
15. Davison, B. Web Traffic Logs: An Imperfect Resource for Evaluation. In Proc. of 9th Annual Conference of the Internet Society (INET '99). San Jose, CA, 1999
16. Mobasher, B., Cooley, R., Srivastava. J. Automatic Personalization Based on Web Usage Mining. Communications of the ACM, Vol. 43, No. 8, pp. 142-151, 2000.
17. Kimball, R. Margy, R. The data warehouse toolkit: the complete guide to dimensional modelling. John Wiley & Sons, 2002, 2nd ed., 464 p. ISBN 0-471-20024-7
18. Robal, T., Kalja, A., Põld, J. Analysing the Web Log to Determine the Efficiency of Web Systems. Proc. of the 7th International Baltic Conference on Databases and Information Systems DB&IS'2006. Communications, Vilnius, Lithuania, 3-6. July 2006, pp. 264 - 275.
19. The Web Robots Pages, http://www.robotstxt.org/wc/robots.html [2006]
20. SpiderHunter, http://www.spiderhunter.com/ [2006]
21. Tan, P-N., Kumar, V. Discovery of Web Robot Sessions based on their Navigational Patterns, Data Mining and Knowledge Discovery, 6(1), 2002, pp. 9–35
22. MySQL AB, http://www.mysql.com [2007]
23. Middleton, S., De Roure, D, Shadbolt, N. Capturing Knowledge of User Preferences: Ontologies in Recommender Systems, Proc. of the 1st Int. Conference on Knowledge Capture, ACM Press, 2001, pp. 100-107