Spiros Sirmakessis (Ed.)

Adaptive and Personalized Semantic Web
Studies in Computational Intelligence, Volume 14

Editor-in-chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series

can be found on our homepage:
springer.com

Vol. 1. Tetsuya Hoya
Artificial Mind System – Kernel Memory
Approach, 2005
ISBN 3-540-26072-2

Vol. 2. Saman K. Halgamuge, Lipo Wang (Eds.)
Computational Intelligence for Modelling
and Prediction, 2005
ISBN 3-540-26071-4

Vol. 3. Božena Kostek
Perception-Based Data Processing in
Acoustics, 2005
ISBN 3-540-25729-2

Vol. 4. Saman K. Halgamuge, Lipo Wang (Eds.)
Classification and Clustering for Knowledge
Discovery, 2005
ISBN 3-540-26073-0

Vol. 5. Da Ruan, Guang Chen, Etienne E.
Kerre, Geert Wets (Eds.)
Intelligent Data Mining, 2005
ISBN 3-540-26256-3

Vol. 6. Tsau Young Lin, Setsuo Ohsumi, Chunn-Jung Liu, Xiaohua Hu, Shusaku
Tsunomo (Eds.)
Foundations of Data Mining and Knowledge
Discovery, 2005
ISBN 3-540-26257-1

Alpaslan, Lakhmi C. Jain, Srikanta Patnaik (Eds.)
Machine Learning and Robot Perception,
2005
ISBN 3-540-25549-X

Vol. 8. Srikanta Patnaik, Lakhmi C. Jain,
Spyros G. Tzafestas, Geramno Resconi,
Amit Konar (Eds.)
Innovations in Robot Mobility and Control,
2006
ISBN 3-540-26892-8

Vol. 9. Tsau Young Lin, Setsuo Ohsumi,
Chunn-Jung Liu, Xiaohua Hu (Eds.)
Foundations and Novel Approaches in Data
Mining, 2005
ISBN 3-540-28315-3

Vol. 10. Andrzej P. Wierzbicki, Yoshiteru
Nakamori
Creative Space, 2005
ISBN 3-540-28458-3

Vol. 11. Antoni Ligęza
Logical Foundations for Rule-Based
Systems, 2006
ISBN 3-540-29117-2

Vol. 13. Nada Nedjah, Ajith Abraham,
Luiza de Macedo Mourelle (Eds.)
Genetic Systems Programming, 2006
ISBN 3-540-29849-5

Adaptive and Personalized Semantic Web,
2006
ISBN 3-540-30605-6

Spiros Sirmakessis
(Ed.)

Adaptive and Personalized Semantic Web
Foreword

Web Personalization can be defined as any set of actions that can tailor the Web experience to a particular user or set of users. To achieve effective personalization, organizations must rely on all available data, including the usage and click-stream data (reflecting user behaviour), the site content, the site structure, domain knowledge, as well as user demographics and profiles. In addition, efficient and intelligent techniques are needed to mine this data for actionable knowledge, and to effectively use the discovered knowledge to enhance the users’ Web experience. These techniques must address important challenges emanating from the size and the heterogeneous nature of the data itself, as well as the dynamic nature of user interactions with the Web. These challenges include the scalability of the personalization solutions, data integration, and successful integration of techniques from machine learning, information retrieval and filtering, databases, agent architectures, knowledge representation, data mining, text mining, statistics, user modelling and human-computer interaction. The Semantic Web adds one more dimension to this. The workshop will focus on the semantic web approach to personalization and adaptation.

The Web has been formed to be an integral part of numerous applications in which a user interacts with a service provider, product sellers, governmental organisations, friends and colleagues. Content and services are available at different sources and places. Hence, Web applications need to combine all available knowledge in order to form personalized, user-friendly, and business-optimal services.

The aim of the International Workshop on Adaptive and Personalized Semantic Web that was held in the Sixteenth ACM Conference on Hypertext and Hypermedia (September 6-9, 2005, Salzburg, Austria) was to bring together researchers and practitioners in the fields of web engineering, adaptive hypermedia, semantic web technologies, knowledge management, information retrieval, user modelling, and other related disciplines which provide enabling technologies for personalization and adaptation on the World Wide Web.
Topics of the Workshop include but are not limited to:

- design, methodologies and architectures of adaptable and adaptive Web information systems
- user interface design for adaptive applications on the Web
- semantic web techniques for adaptation
- authoring of adaptive hypermedia for the Web
- distributed user modelling and adaptation
- semantic web mining
- personalized taxonomies or ontologies
- hybrid recommendation systems
- model integration for personalization and recommendation systems
- web usage, content, and structure mining
- automated techniques for generation and updating of user profiles
- machine learning techniques for information extraction and integration
- applications of relational data mining in personalization
- adaptive personalized web applications

This workshop could not have been held without the outstanding efforts of Marios Katris at the workshop support. I would like to thank the programme committee members for their efforts and support. Finally, recognition and acknowledgement is due to all members of the Internet and Multimedia Research Unit at Research Academic Computer Technology Institute.

Dr Spiros Sirmakessis
Assistant Professor
R.A. Computer Technology Institute
syrma@cti.gr

Programme Committee and Reviewers

Bamshad Mobasher, School of Computer Science, Telecommunication, and Information Systems, DePaul University, USA.

Olfia Nasraoui, Dept of Computer Engineering & Computer Science, the University of Louisville, USA.

Lars Schmidt-Thieme, Computer-based New Media, Institute for Computer Science, University of Freiburg, Germany

Martin Rajman, Center for Global Computing, EPFL, Swiss Federal Institute of Technology, Switzerland.

Ronen Feldman, Department of Computer Science, Bar-Ilan University, Israel.

Spiros Sirmakessis, Computer Technology Institute and Technological Educational Institution of Messolongi, Greece.

Peter Haase, Institute AIFB, University of Karlsruhe, Germany.
An Algorithmic Framework for Adaptive Web Content

Christos Makris, Yannis Panagis, Evangelos Sakkopoulos and Athanasios Tsakalidis
Research Academic Computer Technology Institute, RU5, Patras, Greece
Dept. of Computer Engineering and Informatics, University of Patras, Greece
{makris,sakkopul,panagis,tsak}@ccti.upatras.gr

Abstract. In this work a twofold algorithmic framework for the adaptation of web content to the users' choices is presented. The main merits discussed are a) an optimal offline site adaptation—reorganization approach, which is based on a set of different popularity metrics and, additionally, b) an online personalization mechanism to emerge the most "hot" (popular and recent) site subgraphs in a recommendation list adaptive to the users' individual preferences.

1 Introduction

User driven access to information and services has become more complicated, and can sometimes be tedious for users with different goals, interests, levels of expertise, abilities and preferences. The Boston Consulting Group announced that a full of 28% of online purchasing transactions failed and 1/3 of them stopped shopping on-line due to usability difficulties [2]. This problem is crucial in on-line sales systems with thousands of products of different kinds and/or categories. It is obvious that typical search methods are becoming less favorable as information increases resulting in money losses.

In user-centered applications, two parameters affect usability:

- Orientation and navigation strategy. Users are frequently uncertain as how to reach their goals. Since users have different states of knowledge and experience, information presentation may be too redundant for some of them and too detailed for others.
- Quality of search results. Users cannot locate efficiently the information they need (results must be relevant and come quickly).

Moreover, with the unprecedented growth of the Internet usage, websites are being developed in an uncontrollable, ad-hoc manner, a fact frequently reflected to unpredictable visit patterns. Thus, a critical task for a website
maintainer is to use enumerable metrics in order to identify substructures of the site that are objectively popular.

Web Usage Mining has emerged as a method to assist such a task. The fundamental basis for all mining operations entails processing web server access logfiles. In its most simplified approach, usage mining entails registering absolute page visits or identifying popular paths of information inside a website, by the means of logfile analysis software solutions such as Webtrends (http://www.webtrends.com), and Analog (http://www.analog.cx). When the goal is to detect popular structural website elements, more elaborate techniques have been devised. Some representative work is presented hereafter.

This work contributes two main approaches: it presents an optimal offline site adaptation—reorganization approach based on a set of different popularity metrics and it presents an online personalization mechanism to display the most “hot”—popular and recent—site subgraphs in a recommendation list adaptive to the users’ individual preferences. Both approaches build on well-known results in data structures in the areas of optimal trees and adaptive data structures.

The rest of the paper is organized as follows. Section 2 present background and related work. Section 3 presents popularity metrics that can be taken into account after analyzing user behaviour. Metrics are both localized, i.e. for certain pages and more globalized. Section 4 presents two approaches to reorganize website structure after having computed the appropriate metrics. We conclude with future directions.

2 Background and Related Work

To receive web usage feedback, web sites have been accompanied with logging mechanisms that have been evolving over time. However, the consequences of ad hoc implementation are depicted on the logged navigation trails, where mining for useful information in the logs has become a travel through an information labyrinth.

A shift from standard HTML based applications toward server side programmed web applications is noted several years now, especially with the advent of technologies such as Java servlets, PHP and lately with Microsoft .NET. Among other features, new techniques allow URL re-writing to provide additional information on the HTTP requests, HTML server-side pre-rendering or pre-compilation to facilitate quicker download, client-side code injection to enable transparent capturing of additional user actions and custom logging databases that keep details regarding content delivery to web and proxy servers.

Significant work on converting server logfiles to valuable sources of access patterns has been conducted by Cooley [6]. Apart from analyzing logfiles, it is important to use analysis as input and determine which changes, if any, to bring to the website structure. Chen et al. [3] describe efficient algorithms to infer access patterns corresponding to frequently traversed, website paths. Apart from analysing logfiles, it is important to use analysis as input and determine which changes, if any, to bring to the website structure. Srikant and Yang [10] infer path traversal patterns and use them to indicate structural changes that maximize (or minimize) certain site-dependent criteria. Finally, in [4, 5] the authors define techniques to assess the actual value of webpages and experiment on techniques and mechanisms to reorganize websites.

3 An Overview of Metrics for Webpages and Site Subgraphs

3.1 Access Smells: Absolute, Relative, Spatial and Routed Kinds

Several different metrics have been proposed to calculate the access frequencies from log file processing (see [7] for an early analysis on web logs). In this section, we present access smells, which are different kinds of metrics to estimate a web page’s popularity. We present the approaches of [4] and [8].

As absolute kind of calculation, we refer to the Absolute Accesses (AA) to a specific page i of a site. The relative kind has been initially outlined in [8]. It is defined as:

$$ RA_i = a_i * AA_i $$

That is, the $RA_i$ of page $i$ is a result of the multiplication of $AA_i$ by a coefficient $a_i$. The purpose of $a_i$ is to skew $AA_i$ in a way that better indicates a page’s actual importance. Hence, $a_i$ incorporates topological information, namely page depth within site $d_i$, the number of pages at the same depth $n_i$ and the number of pages within site pointing to it $r_i$. Thus $a_i = d_i + n_i/r_i$. According to [8], the number of hits a page receives, as those are calculated from log file processing, is not a reliable metric to estimate the page’s popularity. Thus this refined metric is proposed, which takes into account structural information. Based on this new notion of popularity, reorganization of certain pages is proposed.

In [4] two more kinds have been introduced towards a similar direction. Acknowledging the importance of reorganization proposals, the aim was to further facilitate the idea of reorganization by introducing two new metrics. The first one takes into account both structural information and the differentiation between users coming from within the website and users coming from other websites, while the second uses a probability model to create a suitable refining factor. Key feature of the new metrics is the higher fidelity on the proposed reorganization proposals. In this direction the authors decompose $AA_i$ into two components, $AA'_{i,m}$ and $AA_{i,out}$ to account for absolute accesses from inside the site and from the outside world. Thereby,

$$ RA_i = a_{i,m} * AA'_{i,m} + a_{i,out} * AA_{i,out} $$
We call these metrics the \textit{spatial} and \textit{routed} kind of page popularity, respectively. \textit{Spatial} kind is based on the topological characteristics of a web site. In particular, these characteristics are implied by the fact that a web page may be accessed using four different ways. Firstly it gets accesses originating from the site (neighboring pages), secondly directly via bookmarks stored at a client browser (indicating that a client prefers this site for his/her reasons), thirdly by incoming links from the outside world and finally by typing directly its URL. By considering accesses from inside the site we obtain $AA'_{i,\text{in}}$. Accounting the last three of the above access methods and with the proper normalization (see [4]) we obtain $AA_{i,\text{out}}$.

In the \textit{routed} kind of calculation as presented in [4], the idea was to increase page relative weight, inversely proportional to its access probability. Considering a site structure as a \textit{directed acyclic graph (DAG)} $G$, with $s$ nodes and $v_i$ denoting page $i$. Suppose that a user starts at the root page $v_r$, looking for an arbitrary site page $v_t$. At each node $v_i$ he makes two kinds of decisions: either he stops browsing or he follows one of the $out(v_i)$ links to pages on the same site. If we consider each decision equiprobable, the probability $p_i$, of each decision is $p_i = (out(v_i) + 1)^{-1}$.

Consider a path $W_j = <v_r, v_1, \ldots, v_t>$, from $v_r$ to $v_t$. Counting the routing probabilities at each step, the probability of ending up to $v_t$ via $W_j$, is simply:

$$P_{t,j} = \prod_{v_i, v \in W} p_i$$

There may be more than one paths leading to $t$, namely $W_1, W_2, \ldots, W_k$. The overall probability of discovering $t$, $D_t$ is:

$$D_t = \sum_{i=1}^{k} P_{t,k}$$

For the example in Fig. 1,

$$D_2 = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{27}$$

Considering page $i$ as target, the higher $D_i$ is the smaller $a_{i,\text{in}}$ shall be, thus we choose $a_{i,\text{in}}$ to be, $a_{i,\text{in}} = 1 - D_i$. We also let $a_{i,\text{out}}$ to be one.

Thus we define $RA_i$ as:

$$RA_i = (1 - D_i) \cdot AA'_{i,\text{in}} + AA_{i,\text{out}}$$

with $AA'_{i,\text{in}}$ and $AA_{i,\text{out}}$ defined as previously.

3.2 Recording User Visits and Hot Subgraphs

The mechanism of [4] scans the website graph in order to keep only the non-intersecting subpaths. Suppose that the site is modelled as a graph $G(V,E)$ kept in an adjacency matrix representation, with matrix $A$.

After the completion of the identification of “Maximum Forward Paths” [1], the website access sequences (paths) are kept. Each path sequence $P_i$ of access frequency count ($F_i$), occurs with count ($P_i \geq \min\{\text{support}\}$). In order to keep only these paths we set to zero the corresponding cells in $A$, thereby pruning the remaining paths. Suppose that there are $p$ frequent paths. An array $C[1 \ldots p]$ is kept, such that $C[i] = \text{count} (P_i)$ and $p$ is the maximum number of paths.

A breadth-first search is initialised. Suppose we visit node $v$. If there is only one outgoing edge from $v$ in the graph remaining after the initial pruning, then there must be a single non-zero element in the $v$-th row of $A$. If this is the case, we delete node $v$ by zeroing out the non-zero entry. There is a small subtlety here; $v$ might be part of a larger path. Therefore, we delete $v$ only if there is no entry but zeros in the $v$-th column of $A$ (an indication that no edge ends at $v$). When deleting node $v$ we add the path frequencies of the paths passing through $v$, to their subpaths. For example suppose that we had a path $P_1 = abfh$ and paths $P_2 = bfh$, $P_3 = bh$. After elimination of $a$ we perform $C[j] = C[j] + C[i]$ and $C[k] = C[k] + C[i]$. The procedure stops when there are no nodes left to be eliminated. The remaining non-zero elements correspond to a subgraph of $G$ with high popularity. For a non-zero element $i$, $C[i]$ keeps a popularity metric for path $i$. Actually what remains after this procedure is popular website paths that have common origins, see Fig. 2.

The above algorithm can gracefully adjust to multiple levels of granularity. After setting the threshold at the first path-elimination step, then after each vertex-elimination, one can define a higher threshold to correspond to more
4 Algorithms for Organizing Web Content According to Mined Access Patterns

In this section we present two approaches to be used as a second step after the initial computation of page importance. The results of the first stage are recorded into an adaptive user profile (outlined in [9]) to achieve personalization effects for registered users that visit web sites of interest. The first of them, the offline, uses computed importance weights, to optimally reorganize the structure of a website so that it minimizes the navigation entropy. The second one, the online, adapts the page presentation after each visit to a certain page.

4.1 Some Preprocessing Steps

In order for the following techniques to operate correctly we require the following preprocessing steps to take place.

Consider a metric $M$ over some object $i$, and let $m_i$ denote the corresponding value. We require $M$ to define a probability distribution. Therefore, we require an initial normalization step to take place, i.e. we set $m_i = m_i / \sum_j m_j$.

Furthermore, the data structures used require their items to be items of an ordered set. In order to satisfy this requirement, we traverse the site structure (in BFS or DFS) and assign each page a unique number. In the case of graph components of Sect. 3.2, we can use their unique path identifier instead.

Therefore, from now on we will assume that we work on a set of website elements, single web pages or website components, each of which has been assigned a unique number and a normalized popularity metric.

4.2 Offline Optimal Reorganization Algorithm

After logfile processing, calculating one of the metrics of Sect. 3 and performing the process described at the previous subsection, we have arrived at the following situation: we have a set $W$ of web elements and a probability distribution on $W$. A probability $p(w_i), w_i \in W$ indicates the popularity of $w_i$. We would like to organize $W$ in order to minimize the access entropy on its elements, i.e. our goal is to minimize

$$E = \sum_{w_i} p(w_i) \cdot l_i$$

where by $l_i$ we denote, the path length leading to element $w_i$. This problem is easily reduced to the problem of finding an optimal binary search tree. The most suitable version of this problem in the context of website reorganization, is the construction of a node-oriented search tree. This implies that we allow every node to contain useful information and not only the tree leaves as is the case in leaf-oriented trees. Hence, the equivalent problem is the following:
Given a set of \( n \) values \( W = w_1 < w_2 \cdots < w_n \) and an access distribution \( D = \{ p(w_1), p(w_2), \ldots, p(w_n) \} \) on its elements find the node-oriented search tree that minimizes \( E \).

This problem has been addressed by Knuth [10]. The problem is solved by the means of a modified dynamic programming technique. Straightforward dynamic programming results in a solution with \( O(n^3) \) time. However, after constraining an internal condition at the dynamic programming loop the time complexity drops to \( O(n^2) \) in \( O(n^2) \) space. For more details also refer to Mehlhorn [11], Sect. 3.4.

In order to automatize link creation after construction of an optimal tree structure, the framework introduced in [5] can be of great assistance.

We can extend the previous solution by embedding splay tree ([12]) heuristics while maintaining and reorganizing the structure of the Web Site. A splay tree is a binary search tree in which all operations are performed by means of a primitive, the so-called splaying operation. A splay operation on a node \( x \) consists of a sequence of rotations from the node \( x \) to the root of the tree with each rotation rotating some higher edges before some lower edge and the whole sequence of a rotations corresponding to a splay leads to decreasing the depth of the nodes in a path by about a half. A crucial property that splay trees have is that they are statically optimal.

In order to apply splay trees in our setting we have firstly to transform the tree structure of the Web Site into a binary tree which is easy to do by replacing each node with degree \( k \) by a \( \log k \) height binary tree. Hence each node of this tree has a weight corresponding to each importance metrics and the weight of each node is updated at a rate that is settled by the Web Site Administrator. Whenever the importance metric of a node is updated then the whole path from the node to the root is updated and so changes to the Web Site design can take place in the path from the leaf to the root. Details of how to implement this strategy are omitted from this short version of the paper.

4.3 Online Personalization Using Adaptive Data Structures

The previous approach was static in the sense that access results are gathered after a period of observation, access metrics are computed and then restructuring is performed. In many cases though, it is preferable to adapt content online and, e.g. give user a link table containing the most frequently and most recently accessed web site parts. A simple and elegant strategy to achieve this goal, without even the need to know the specific popularity of certain web elements, is to use an adaptive data structure. In the following we constrain for the sake of clarity our discussion to web pages.

The data structure that can be used is the adaptive list. The adaptive list is a doubly-connected list of unordered items. Each time an item is accessed, it is brought to the front (left end) of the list. This adaptation rule is called Move-to-Front. An example is shown in Fig. 5.

\[
2, 7, 5, 1, 4, 3, 6 \rightarrow Access (3) \rightarrow 3, 2, 7, 5, 1, 4, 6 \rightarrow Access (7) \rightarrow 7, 3, 2, 5, 1, 4, 6
\]

Fig. 5. The Move-To-Front rule

It is proved (see e.g. [11]) that Move-to-Front is at least 2-competitive, i.e. the total running time for a sequence of element accesses is at most twice slower than the optimum adaptation strategy. Note that the optimum strategy has full knowledge of the element distribution, whereas the adaptive list achieves its competitiveness without any knowledge of the distribution.

The adaptive list and the Move-To-Front rule can be the structure of choice whenever we want to keep a recommendation list of the most recently visited pages. These lists are kept for each web page and for each user in a personalized web environment and present users with possible navigation choices. In a possible implementation we can present users the leftmost \( k \) elements of the list, where \( k \) is a predefined constant. This amounts to presenting user with the \( k \) pages that she is most likely to visit in the future.

5 Conclusions and Future Steps

In this paper we have proposed two different approaches to adapting online content. Both our approaches are based on established results from the data structures’ area and have aimed to provide a new viewpoint to a classical concern in online content. Those approaches are both elegant and easy to implement. Future steps include the description of a framework that it would evaluate the combination of reorganization metrics with different sets of redesign proposals. We also consider as open issue the definition of an overall website grading method that would quantify the quality and visits of a given site before and after reorganization, justifying thus the instantiation of certain redesign approaches.

Acknowledgments

The work of Yannis Panagis and Evangelos Sakopoulos is partially supported by the Caratheodory research programme at the University of Patras. Yannis Panagis is also like to thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II) and particularly the Program PYTHAGORAS, for funding part of the above work.
References