Acquisition and Modelling of Short-Term User Behaviour on the Web: A Survey

Ondrej Kassak, Michal Kompan and Maria Bielikova

Faculty of Informatics and Information Technologies,
Slovak University of Technology, Ilkovicova 2,
Bratislava, 841 04, Slovakia
E-mail: ondrej.kassak@stuba.sk; michal.kompan@stuba.sk;
maria.bielikova@stuba.sk

Received 5 February 2018; Accepted 10 August 2018;
Publication 17 September 2018

Abstract

User behaviour in data intensive applications such as the Web-based applications represents a complex set of actions influenced by plenty of factors. Thanks to this complexity, it is extremely hard for human to be able to understand all its aspects. Despite of this, by observing user actions from multiple views, we are able to extract and to model typical behaviour and its deviations on the Web. The website itself, together with transaction server logs, includes information about the site structure, content and about the actual user actions (clicks) within the site. User actions logically reflect the behaviour, while other sources indicate his/her context. Combination of these data sources allows to model the typical user behaviour and his/her preferences. The long-term behaviour describes relatively stable user preferences based on extensive user history. As the Web has become more and more dynamic, modelling user behaviour from the long-term perspective does not satisfy requirements of current Web based applications. On the other side, the short-term behaviour describes current user activity and his/her actual intent. However, this source of information is often noisy. To address these shortcomings the state-of-the-art combines both perspectives, which allows to meaningful and timely modelling of user behaviour. In this paper, we provide a comprehensive survey

Journal of Web Engineering, Vol. 17_5, 23–70.
doi: 10.13052/jwe1540-9589.1752

of user modelling techniques. We analyse types of data sources used for the modelling and approaches for its acquisition. Additionally, we discuss approaches considering actual trends of dynamically changing websites. This trend brings new challenges, which have to be addressed in design and implementation of novel Web applications.

Keywords: User modelling, Short-term user behaviour, Session, User preference, Web-site mining, Usage data mining.

1 Introduction

As the Web became an essential part of our daily lives, more and more activities and time is spent on browsing by the users. At a first sight it seems to be a random sequence of actions, but the opposite is usually true. According to actual user context including personal stimuli, his/her activity is influenced by multiple factors. These factors cover user preferences, personal characteristics (e.g., age, education, previous knowledge), actual information needs (e.g., search for information, browsing the news) and also actual context (e.g., actual trends, events, stimuli from other users) [81]. Moreover, the website itself (e.g., a site structure, its content and a frequency of content changes) influences user behaviour, respectively.

In general, the process of identifying and maintaining user preferences and his/her behaviour is known as user modelling. User modelling on the Web is essential for various tasks as personalization, prediction of user's next behaviour, his/her intent to leave the site, buy some product and many more. Moreover it influences the design and implementation of Web application itself. Often these are entirely based on the ability to model user behaviour. Based on the user model application, the user behaviour is typically captured from the long- and/or short-term perspective. The long-term perspective covers mostly user's typical and stable behaviour. It offers relevant information about user preferences but demands high amount of data about user activity (to create a "good" user model). Another shortcoming is long reaction time to change in user preferences, which is crucial mainly for dynamic sites. Such a model is typically used for recommendation of interesting content based on stable interests or site adaptation based on typical user's behavioural patterns. The short-term user model, on the other hand, covers actual behaviour, so it is capable to react to recent impulses. In comparison to the long-term model, it fails modelling all relevant information (i.e., it covers shorter time periods and thus less user actions are used).

Obviously, there are domains with different characteristics considering frequency of change, e.g., there are highly dynamic domains such as news or multimedia; average dynamic domains such as e-commerce or low dynamic domains such as personal sites. Dynamic domains are challenging from the user behaviour analysis and modelling point of view. In other words, it is difficult to acquire and maintain user preferences due to the frequent changes in the content [145]. Moreover, as the Web is characterised by plenty of anonymous or occasional users, whose previous preferences are unknown [133], the short-term user behaviour modelling is actually gaining importance. It helps to improve user experience by providing adaptive or personalised services able to react to the latest user behaviour in the real time.

From the perspective of the short-term user behaviour modelling, we recognise two basic information sources (Figure 1): website data mining which is based on the structure and content of its webpages; and the usage data mining based on the users' action acquisition and session modelling. In traditional long-term behaviour, the division of data sources is similar, but their importance quite differs. The short-term modelling emphasises the usage data (it is important which specific action user performed). In the long-term perspective we model preferences on the higher level (e.g., which topics the user visits repeatedly, what knowledge level of some concept he/she has).

The aim of this paper is to offer an overview of the user modelling from the data acquisition phase to the actual building of the model. Existing works dealing with user modelling focus mainly on the phase of the user

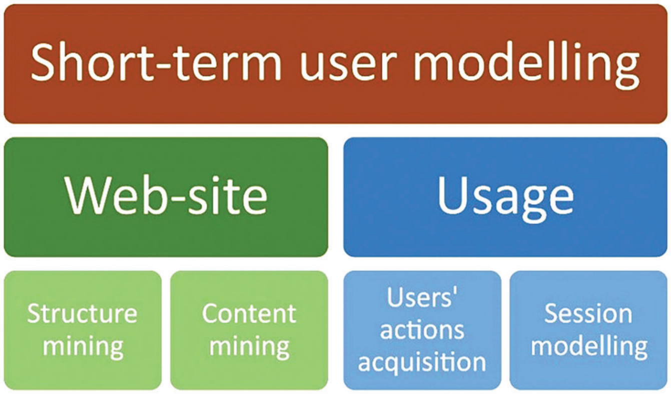


Figure 1 The components of short-term user behaviour modelling. Generally, the Web mining is a subcategory of the standard data mining techniques, while specific website characteristics are taken into account. For this reason, only a subset of data mining techniques is used [7].

model application. The reason is that the user model quality is evaluated primarily indirectly by the mean of its application for a particular task (e.g., recommendation). This pushes the model into the position of a source of attributes for the chosen application task. Previous surveys on user modelling focus mainly on the long-term behaviour, which is more stable and easier to obtain, but it does not fully cover the actual user actions [1]. On the contrary, in [95] authors deal with both, the long- as well as short-term behaviour. Their survey is however limited only to specific domain of social annotations. In our paper, we offer a complex overview of usage of short-term user modelling as a whole, with respect to a great variety of website domains.

The paper is divided as follows. In Sections 2 and 3 we focus on data acquisition for user modelling, namely we describe data mined from the website structure and content (Section 2), and the usage data (Section 3). The usage data represents the most important source of information for the short-term user modelling as it describes the user behaviour directly. The process of user modelling is described in Section 4. We focus on user model representations, the process of user model initialisation, and methods of preference modelling and maintaining in highly dynamic domains.

Highlights of the survey are as follows:

- Overview of website mining focused on site structure and page content
- Comparison of usage mining approaches focused on actions acquisition and session modelling
- Comprehensive analysis of user modelling focused on multi-layer time modelling in dynamic domains

2 Website Data Mining

The user behaviour on the website is influenced by many factors (e.g., actual aim, context, previous knowledge), which together form user decisions (e.g., how long to stay on the actual page, which page to visit next). To be able to analyse the user behaviour, it is necessary to obtain data and information about the actions that the user accomplished in the past, his/her characteristics and specific features of the site itself.

Authors in [78] described the process of Web mining, which they divided into three parts, based on the area of their focus:

- *Website structure mining* – acquiring and processing data about the website as a whole, its structure, topology and model underlying the link structure.

- *Website content mining* – acquiring and processing data about individual pages, their description, content and the information architecture of the site.
- *Website usage mining* – acquiring and processing information about the user actions on the site, mapping them into sessions, recognising session authors and identifying the intent.

Each part helps us to reveal additional information, which together improve the user behaviour modelling.

2.1 Website Structure Mining

The aim of the website structure mining is primarily to discover and to understand a structure of the site. Secondly, to create a model of site pages and their mutual hyperlink connections. In this case, the content of individual pages is ignored, only the topology is considered. As users browse mostly through directly connected pages within the site, the site model represents a set of most probable visit paths [114].

2.1.1 Website Structure Minig

Understanding the website structure allows us to find hidden connections, which may affect the user behaviour. In early studies, the structure of the website was represented as a multidigraph[1] G [17]:

$$\begin{aligned} G = (V, E); V = \{v_{Entrance}, v_{Exit}, v_1, v_2, \ldots, v_n\}; \\ E = \{v_i v_j, v_i v_k, \ldots, v_l v_m\} \end{aligned} \quad (1)$$

where vertexes V represent the website pages, which are identified by unique URL addresses and described by their content. The edges E connecting the pairs of vertexes, represent the hyperlinks between the pages. In addition, the graph contains two more special vertexes that represent an entrance $v_{Entrance}$ and an exit v_{Exit} of the site. These vertexes represent the rest of the Web. Both of special vertexes are connected with every regular vertex in the graph as users can visit and leave the modelled site in every one of its pages [27].

This old representation was proposed mostly for static pages that did not change a lot. This is a major disadvantage – it requires pages to have a static content, which does not change over the time (the model is unable to model an update of page content).

[1] A graph containing oriented edges, multiple edges between the same pair of vertexes might occur [17].

Nowadays, the Web mostly contains dynamic pages, which are updated in short time intervals. Often the content is adapted or personalised for specific users [2]. This trend is domain dependent and it is stronger for news, events or personal websites than for government or educational websites. In other words, one of the dynamic sources is the user generated content. Also, according to [2], the pages with deep URL structure change less often than root pages. Their changes are, however, more fundamental in comparison to the root page changes (typically by adding or removing links to pages with deep URL structure).

Another problem of this representation is that it does not consider the *importance* of pages (vertexes). The solution is to use ranking algorithms, which also calculate page importance for dynamically changing sites, e.g., HITS [76], PageRank [18] or weighted PageRank [100].

Kleinberg [76] in his HITS algorithm represents the website as a graph and its pages as vertexes of two types – authorities and hubs. Authorities are globally important or contain interesting content. The page is an authority when there are lot of pages pointing to it. Hubs serve as catalogues. They, alone, do not contain an important content, their power is in referencing the high number of authority pages. In addition, [18] in the PageRank algorithm used the idea of calculating the rank of the page based on the number and the quality of hyperlink connections pointing to this page. In comparison to the HITS algorithm which considers both inbound and outbound hyperlink connections, PageRank calculates the score based only on inbound links.

The page importance is beneficial for the task of the user behaviour analysis or prediction. It helps us to understand specific behaviour as for example a fast clicks sequence – as the user probably accessed the authority page through a hub page etc. Thanks to the iterative computation, both ranking algorithms are resistant to site updates and dynamic page changes [100].

This is helpful for websites with high update rate, i.e., week [107] or even day interval [34]. In 2004, Ntoulnas et al. [107] report mostly static websites in their experiment. The content of 65% website pages does not change and if it even does, mostly minor changes were observed. Few years later, however, Adar et al. [2] concluded that on dataset of 55 000 pages, 41.6% of them changed within one hour. This trend was also supported in [55] where authors claimed that the rate of pages moving from static to the dynamic content is close to 100%.

Clearly, a good indicator of the future page change rate is the rate of its past changes [2]. An extensive study performed for 3 months over 151 mil. of pages found that there was 22% of pages deleted and 34.8% of pages updated. Larger pages were updated more often [48].

As the Web became more dynamic, there is also a need to consider that the page can lose its position of the fundamental source of information in its description of the overall Web structure (e.g., single page applications). For this reason, it is more suitable to represent pages by URLs with their query parameters, because this step helps to preserve the uniqueness of the site vertexes [9]. The idea of describing the website structure by vertexes with rich metadata description is supported also in [83]. Lee et al. however joined these vertexes into the tree instead of the graph.

The *temporality* of pages was studied by Desikan and Srivastava [41], who proposed a set of multiple level characteristics to describe changes made on site over the time. These characteristics help to explain changes made in the selected time period:

- *Temporal "single node" characteristics* – describe properties of pages in the time, when their content was not changed, e.g., the frequency of page access, the level of cluster change over the time.
- *Temporal "sub graphs" characteristics* – describe properties of pages, when there were only minor changes made to the site, e.g., the order of pages, site size, PageRank value, max. authorities.
- *"Whole graph" characteristics* – a set of characteristics describing the website graph properties, e.g., basic (order, size) or derived ones (max. hub score, average hub score used in PageRank algorithm [110]).

Nowadays, the complexity of websites structure has been rising exponentially. Despite this fact, large scale analysis of hyperlinks showed that there exists a Power Law distribution $p(x)$ [8]:

$$p(x) = x^{-\alpha} \tag{2}$$

For various site characteristics, this distribution of some argument x differs only in value of the exponent α [8]:

- The number of pages per website describes the distribution with $\alpha \in \langle 1.2, 1.6 \rangle$ [42]
- The number of hyperlinks pointing to the page is close to the distribution with $\alpha \in \langle 1.6, 2.1 \rangle$
- Piecewise power law with $\alpha_1 \in \langle 0.3, 0.7 \rangle$ and $\alpha_2 \in \langle 1.9, 3.6 \rangle$ describes the number of hyperlinks outgoing from a page
- $\alpha \in \langle 1.8, 1.9 \rangle$ represent the distribution for the ranking number in the PageRank algorithm

In addition, the website can be described by other metrics expressing its structure as for example *coverability* and *reachability* (both describing the difficulty of visiting one site page from another) proposed in [129]. In this approach, the website is modelled as a graph, where pages are represented as vertexes and hyperlink connections as the edges. Next, an update policy has to be chosen, while it should consider a frequency of the page content change.

To sum it up, the structure of the website represents an important source of information for the task of user modelling. It prejudices possible user actions (page visits) and influences the sequence in which user will probably perform them. The website structure also allows to identify an importance of individual pages (based only on hyperlink connections between them) by several ranking algorithms. The structure of the website is typically represented by the graph or tree [83], where vertexes represent pages and edges the hyperlink connections. Nowadays, this representation is hampered by frequent website updates for many domains. The typical solution is to set a regular update policy. This, however, does not cover all complications caused by website updates (e.g., it does not describe updates of the page content). Various authors address this problem by construction of multiple website models over time (memory ineffective) or by a modelling a new vertex in case of the page content update (increases model complexity).

2.2 Website Content Mining

The content is the main reason why a user visits the website and its pages. Its quality affects the visit duration and future user return. The knowledge about the content, which user experienced in the past and which topics he/she preferred, helps to predict user's future steps, recommend him/her interesting content or generally improve his/her experience. Based on page content, there is possible to estimate its importance for the user. For these reasons the step of website content mining plays important role in the website mining and thus user behaviour modelling process.

The page is typically composed of semi structured data. This data contains several structured elements (e.g., HTML elements) comprising the unstructured text in natural language. To be able to extract the page content, it is useful to segment these structured elements and extract the information from those containing the main page content. Nowadays, several approaches are used to extract relevant website information.

Supervised machine learning approaches need generally large number of elements manually labelled by a domain expert. This process is time consuming, due to the large number of different website and page types. On the contrary, nowadays automatic pattern discovery (unsupervised machine learning) eliminates the need for manual pre-processing, but the precision usually drops and a high number of assumptions is produced [47].

As a reaction to these (supervised and unsupervised), a third type was proposed. Identification of page elements by a method of visual information used to segment data. This approach also facilitates aligning and extracting the content from identified elements by partial alignment based on a tree matching. This approach produces accurate alignment of various data types. Experimental results using large number of pages from diverse domains showed that this approach is able to segment, align and extract data very accurately (approx. 96% precision) [47]. The concept of a website content structure extraction based on a visual site segmentation was introduced by Cai et al. in VIPS algorithm. The main advantage of this approach is that it extracts page segments at a semantic level. In this way, the approach works well even when the physical page DOM structure is far different from its visual presentation [23].

Finally, the natural language and text processing approaches should be applied on extracted page content. From the machine processing and further modelling purposes we are usually interested in the extraction of highly informative words [25]. Some kind of semantics is often extracted, e.g., Named entities [137], special phrases [119]. Nowadays, there is also a trend of combining different data mining algorithms (i.e., ensemble learning), which are together able to mine more complex results from the given source [93]. In the case of short text containing only limited volume of data (e.g., social media posts), there are approaches mining additional data from external sources based on key phrases mined from original source [72].

To understand the user behaviour on the website, at first, it is crucial to understand the content he/she interacts with. In this section we already shown the approaches to mine the segments with the content from the pages. The most important, however, is to understand the content keywords and/or semantics. For this reason, following subsections show techniques for website content modelling and information architecture identification.

2.2.1 Website Content Modelling

From the statistical point of view, there are regularities in the text distribution within the pages. Ipeirotis and Gravano [67] pointed that ranked number of words used in a single page, subjects to the Zipf's Law distribution with the power of $\alpha \cong 1$. This shows that for the further computation processing, the extracted text has to be modelled in the means of reducing number of words by representing content by some features (e.g., latent).

The text processing is an extensively studied research area, which aims at extraction distinctive text features. The statistics based approaches are often used as a first choice thanks to their simplicity. For better results, the usage of words semantics seems to be a a promising idea in next years.

Notoriously known Term Frequency-Inverse Document Frequency model represent most used statistic approach, which considers term importance and its occurrence in the specific document [70]. Gao et al. [49] proposed a TF-IDF extension, which considers calculation of time sensitive frequencies. The TextRank algorithm in contrast to TF-IDF does not requires a corpus, because the importance of keywords is calculated based on its neighbours' frequency and its distance [96]. The Likey method uses the reference corpus, but it calculates keywords importance based on N-gram frequencies [112].

Statistics based approaches are based on word occurrence frequencies. As they use only raw text, their results produce quite similar results, while there is no clear winner. For more sophisticated tasks, as keywords extraction or its organising into hierarchies, unsupervised machine learning approaches are used (e.g., clustering) [134]. Thanks to the various clustering models – centroid, distribution, density or connectivity based [13] – also hidden relations can be found in high dimensional text document.

The step forward represents the semantics in the text processing and modelling. The natural language processing (NLP) approaches are helpful for multiple tasks, e.g., part of speech tagging, named entity recognition or sentiment analysis [5]. Another popular method used to enrich information extracted from the page content is Latent Dirichlet allocation (LDA). In this approach, the pages are described as a mixture of topics. In this way, the page could be described by a set of latent topics that improve similarity search within the site pages and user interests modelling [15].

An extension of standard LDA represents the Online Latent Dirichlet Allocation (OLDA) which automatically recognises thematic patterns, their changes over time and incrementally builds a model (single data pass). Its advantage is the ability to incrementally update the model using Empirical Bayes method according to the information inferred from the new stream. It also provides an efficient way to dynamical tracking and detection of topics in online time [4]. Another LDA extension proposed by Li et al., who came up with the group of LDA algorithm identifying global as well as local topics based on the similar document clustering. This brings a possibility to model topics for whole sites as well as for specific sections [85].

To capture the content evolution of the large websites in the time, there is often used a family of probabilistic time series models [16]. They are used to analyse and to create the space of state models representing topics. The idea is based on assigning preference measures to the categories. The set of categories, however, is valid only for a limited time period (e.g., year). After elapsing this time period, new categories reflecting actual state are formed. In this way, new trends or extensions in domain are reflected, which guarantees the freshness of domain model [16].

In recent years, the research has also been focusing on distributed text representations, which uses embedded models to identify important text features by eliminating less crucial and redundant pieces of information [92]. This allows to model representation optimising the trade-off between training performance and domain portability [60]. The next step is a usage of neural networks, which are nowadays, in context of content modelling, used for distributed words tagging, chunking, Named entity recognition, word representations surpassing [132] or automatic text representation [50].

The selection of specific algorithm should always consider actual usage needs and resources available. In the case, that there is needed a quick result over relatively small data (pages) collection, the simple statistical approach is sufficient. On the contrary, extracting semantic information or representation trained over massive data collection refers to advanced, computationally demanding approaches. Using one approach or another, the information about site content represents a valuable source of information for user behaviour modelling process.

2.2.2 Website Information Architecture Identification

The website information architecture describes the organisation of page elements, their readability, labelling, quality and also ease of search or navigation, which may help us in the user behaviour modelling [103].

After a user visits the website, he/she makes a subconscious decision if this page addresses his/her information needs or he/she has to go to a different website [103]. From the website perspective, it is very difficult to attract a new user and even more difficult to persuade him/her to return in the future (depending on a specific domain). According to Morville and Rosenfeld [103], user typically does not care about the information architecture of the website, but he/she intuitively perceives it by the ability to solve his/her actual task on the page quickly, easy and without high cognitive effort. If the user is not able to address the information need intuitively, he/she will most probably leave the page. Authors divide the information architecture components of the website into these four categories:

- *Organisation* – site components belonging to this category contain major website information, typically divided into several areas according to their meaning or content. They represent the website information backbone and their aim is to transfer the information to the user. Various criteria are used as the topic which they describe (e.g., products offer, company description), relevancy or time (e.g., most read, newest).
- *Navigation* – navigation website components help the users to move through the website, e.g., menus, navigation marks, filters or links to popular or recently read content.
- *Search* – website components used to improve accessibility by the direct search. To this category belong also specific search components as in the calendar, a favourite content, etc.
- *Labelling* – website components (e.g., sections, links) are more usable for the users when they are labelled in language that is meaningful to users (e.g., the menu item labelled as "about" which is more describing as the label "info").

According to Morville and Rosenfeld [103], the quality of the website information architecture is measured by users' ability to find the answers easily for the following questions concerning described categories:

1. *Where am I?* This informational question aims at answering the basic information about the website (e.g., name, motto and logo). It belongs to organisation information architecture category.
2. *I know what I'm looking for; how do I search for it?* The aim of this question is to judge how fast and is the user able to find some exact information on the website without repeated browsing. This question characterises the search category.
3. *How do I get around this site?* The user often does not know properly what he/she is looking for (e.g., recipe for Sunday lunch) but he/she suspects at least the topic. This question is used to test the navigational category.
4. *What's important and unique about this organisation?* To be able to trust the website or to spend there some resources (e.g., money in e-shop, time on blog), user should have access to website background information. This question belongs to the labelling and also to the navigational category.
5. *What's available on this site?* User often visits page from an external link and he/she does not know the page content. Aim of this question is to test how interesting information the page offers to the user in first moments

of the visit. It should contain fresh news and relevant information which should be easy accessible (not to overload the user). This task represents the organisational category.

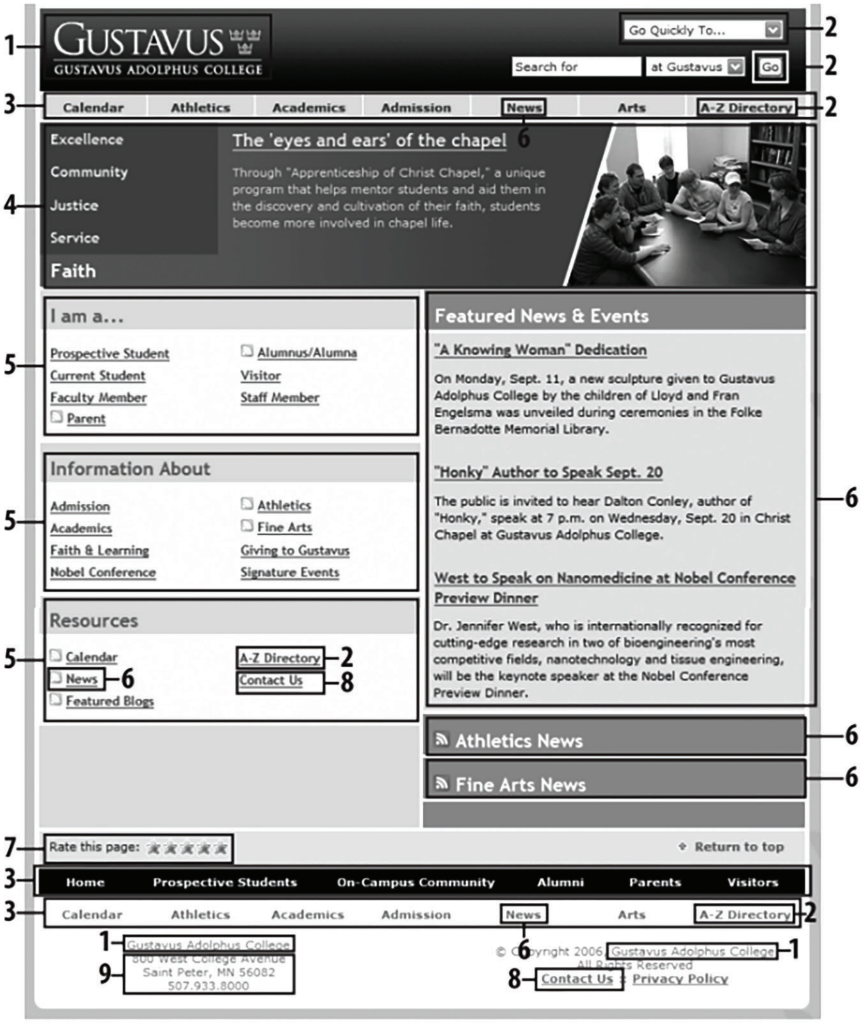
6. *What's happening there?* Information about ongoing activities contributes to the user's good awareness and it is important for him/her to quick introduction. This question belongs to the organisational category.
7. *Do they want my opinion about their site?* If user has the possibility to offer feedback easily in the website, it gives him/her the feeling that the site authors care about his/her opinion. This truly increases the user's positive opinion to the website. This question represents organisational category.
8. *How can I contact a human?* If a user did not find some information from the website or if he/she wants to get some additional one, there should be provided the contact information.
9. *What's their address?* Completion of questions 4 and 8 with additional specialisation.

Based on these questions we can qualitatively assess the information architecture of the website. It is based mostly on optimal combination and the position of the page components. The Figure 2 shows an example of the page with labelled components and the question numbers they answer.

An additional information about the website content can be extracted from its information architecture. As we shown, the website architecture express the effort for users to solve corresponding tasks. Four categories of information architecture together describe the site quality and influence the user experience. In recent years, automatised approaches for architecture design were proposed [116]. This results in more logically organised pages, which reduces users confusion and the chance of leave the site prematurely. As the information architecture is key part of the website content and also structure, it helps to understand the quality of content and thus also user decisions.

3 Website Usage Mining

The user actions, clearly, represent one of the most important sources describing the user behaviour and preferences. These actions are typically acquired reactively from simple server page logs. In the opposite, proactive approaches bring more sophisticated actions collecting by specialised tracking applications [66].



The 'eyes and ears' of the chapel

Through "Apprenticeship of Christ Chapel," a unique program that helps mentor students and aid them in the discovery and cultivation of their faith, students become more involved in chapel life.

"A Knowing Woman" Dedication

On Monday, Sept. 11, a new sculpture given to Gustavus Adolphus College by the children of Lloyd and Fran Engelsma was unveiled during ceremonies in the Folke Bernadotte Memorial Library.

"Honky" Author to Speak Sept. 20

The public is invited to hear Dalton Conley, author of "Honky," speak at 7 p.m. on Wednesday, Sept. 20 in Christ Chapel at Gustavus Adolphus College.

West to Speak on Nanomedicine at Nobel Conference Preview Dinner

Dr. Jennifer West, who is internationally recognized for cutting-edge research in two of bioengineering's most competitive fields, nanotechnology and tissue engineering, will be the keynote speaker at the Nobel Conference Preview Dinner.

Gustavus Adolphus College
800 West College Avenue
Saint Peter, MN 56082
507.933.8000

© Copyright 2006 Gustavus Adolphus College
All Rights Reserved

Figure 2 An example of the webpage with highlighted components describing information architecture (marked with numbers referring to category questions) [103].

The individual user actions are, however, insufficient as direct input for behaviour modelling, because it is difficult to impute complex information directly from simple actions. For this reason, the usage mining process consists of several steps.

Firstly, there is important to focus on the identification of related actions and their grouping into sessions. Authors in [125] defined the session as a trajectory of pages visited by the user sequentially in the time. In the context of the usage mining, a session represents a user activity in the website or websites where the user performs one task (which we may formulate as addressing his/her information need) [124].

The actions performed within the same task typically relate together as they were often realised under the same context and conditions. Thus, actions from one session contain similar characteristics. As the sessions offer more information than their actions individually, it is beneficial to model user behaviour on the level of sessions rather than simple actions.

Actions grouped into one session have to be wisely represented (in order to mine relevant information). Several approaches have been proposed, e.g., per page weight, graph representation or representation considering text weighting [114].

One of the most advanced information to be mined is the identification of actual user intent within the session. According to Broder [19], three types of session intent exist, which are based on user Web information needs: the informational, navigational and transactional sessions. They help us in the user behaviour modelling (by considering user and site context).

3.1 User Actions Acquisition

Two types of user actions are distinguished within the website. Firstly, we have actions that user performs directly on the website (obtained from the page visit logs). The second type describes user behaviour indirectly, i.e., based on the combination of behaviour and context description (e.g., intensity of the keystroke, mouse activity, biometric information) [35, 80].

Thanks to the wide availability, the direct page visit actions are more traditional and more frequently used. Its main shortcoming is the bias, which is usual for the data stream of page visits (e.g., we do not know the reason why user did not click on some link) [69]. To overcome this shortcoming additional actions types may help us (e.g., eye tracking, describing user's gaze) [54, 122]. The user's gaze describes the user behaviour more precisely than click stream data, however, it requires expensive hardware, which is not commonly available to users. Moreover, the privacy concerns of some users prevent from wide application [101].

The parallelism of today's Web browsing (parallel browsing) influences user behaviour and is important to consider in the actions acquisition process [82]. Despite the parallel browsing, an active time spent on pages has the Heavy tail distribution, where the major fraction of the visits last only for short time (typically few seconds) [64, 65]. It shows that not all visited pages are equally important for the user and this should be considered in the user behaviour modelling. Additionally, the page visit length depends on the number of hyperlinks on that page and similarly the number of words on page [108].

Generally, user actions are acquired by *reactive approach*. This approach acquires user actions from data logs created by the general website server logger component. As these logs are not specialised for the website usage mining purposes, it suffers from several complications, e.g., insufficient information logged about the actions. On the contrary, reactive approach does not need any special components and is applicable on almost every website without any additional resources [66].

The second – the *proactive approach* – represents the group of methods specialised for user activity acquisition. It provides more information about user actions than the reactive methods. Proactive methods, however, need some specialised tracking software or user logging (e.g., cookies or login based access).

To be able to use acquired user actions for user modelling or more generally for some data mining task, we need to pre-process them. The pre-processing includes discarding invalid site visit logs or non human user logs (automatised robots, worm and hackers' attacks generating high number of visit requests), replace missing values or outliers as we describe in following sections.

3.1.1 Reactive Methods

In general, reactive methods are an indirect way of obtaining the anonymous user activity. This data is usually extracted from transaction server logs, which were not logged specifically for website usage mining purposes. Reactive logs typically contain the identifier of the page, the timestamp of the visit and sometimes some information about the user. The user is typically described by a browser fingerprint or IP address. The problem is, that this information is not unique [111]. That's the reason why reactive usage actions are less valuable.

The negative aspect is a problem to clearly recognise page views of several users (only the heuristic estimation). The reason is that if multiple users with identical browser fingerprint and IP address access the site in the similar time, there is no mechanism to distinguish between them. Several users can share the same fingerprints (identical browsers) and IP address [111].

Another disadvantage is caused by the fact, that the browsing activity is logged on the server side, thus no client side actions are recorded (e.g., the visiting page by forward or backward button). This produces sessions with incomplete hyperlink paths.

On the contrary, the positive aspect of such unambiguous identification is the users' privacy [44, 94]. The reactive methods of user activity data acquisition are used more often than the proactive [125]. The reason is that they do not need any special software or user activity to log the data. For logged users, this approach captures the data of sufficient quality. For ambiguous user identifiers, the disambiguation has to be performed in the session reconstruction process.

3.1.2 Proactive Methods

Proactive methods are designed in respect to further data processing and mining. For this reason, these methods acquire rich data with additional information describing the user (in the data collecting phase). Some of proactive methods are, however, considered as too invasive and are regulated or prohibited by the law in several countries due to the user privacy issues [125]. Spiliopoulou et al. recognise these categories:

- *Cookie based* – Cookies are typically managed by Web browsers and they represent a persistent data structures created and managed by embedded program on the page. They are used to define the user identification by storing a unique identifier. As cookies allow to map users and their activity considered as a private information, users often disable the tracking (due to security and privacy issues) [22].
- *Tracking* – The user activity is tracked via a client application installed on user's device [118]. This tracking approach is forbidden by the law in some countries due to the privacy issues. Despite this, some tracking applications are used to acquire low level events (e.g., mouse clicks, keyboard inputs) [3]. Similarly, these applications are used to track and synchronise multiple signals as the eye movement, mouse clicks or web-camera outputs [91].
- *Logging in* – One of the most reliable way how to obtain the relevant session information is to require logging in. It is one of the simplest approaches, but in practice it is quite difficult to persuade users to register and log in every time they visit the site.

The most effective approach to the user identification is to motivate users to log in (typically quite challenging). Users are, however, typically passive. For this purpose, the tracing applications are used, but they need to be installed to users' devices. The cookie based approaches represent very simple and effective way of the user identification and thus are widely used. They offer a user identification as well as logging and are easy to use for website provider.

3.1.3 Actions Pre-processing

Raw user actions, have to be pre-processed (e.g., discard the invalid logs or robot actions) before the further usage. Patel and Parmar [111] describe three steps of pre-processing:

- *Deleting useless attributes* from logged actions, which do not describe user behaviour. Typically, attributes such as user identifier (e.g., cookie or IP address according to logging method), URL of visited page and time of access are needed.
- *Omitting the irrelevant or redundant information* logged about the page content (e.g., multimedia files, structured HTML elements or page CSS formatting).
- *Removal of failed requests*, not page request actions, HTTP errors etc. These are identified from server POST and GET responses.

Important task in the pre-processing is the identification and filtering out non human users such as crawlers, robots, but also hacker attacks. Luckily, we can identify them by the high number of requests to the site server, realised in short time intervals and/or for a long time period [109].

Hand by hand with non human users filtering, omitting of real user outliers have to be performed. This is generally realised by omitting actions of some percentile of the users with the most actions made (e.g., by the Mahalanobis distance excluding one percentile tail of most active users [117]) or by biologically inspired algorithms [6].

An importance of usage data in the process of user behaviour modelling is obvious. The effort put in the collecting and mining phase results in qualitatively better descriptive attributes for the behaviour modelling.

The reactive methods are simpler and widely widespread as they do not need any specialised software (they process common transactional server logs). On the contrary, server logs are not specialised for purposes of website usage mining, which can limit their usage.

The proactive methods presume the later usage of user activity logs, so they capture rich information. More extensive usage of proactive methods is also limited due to the legal aspect of intensive logging software.

A specific selection of actions acquisition approach should always consider the information value included in logged data with the value that user gets back from the website.

3.2 Session Modelling

Acquired and pre-processed user actions allow us to proceed in the process of capturing user behaviour on the website and to model the sessions. In this step, the actions are joined based on mutual relations, topic or context in which they were made. The assumption is that related actions share common intent and thus should be processed together. A session modelling consists of two phases – a session reconstruction and representation.

The session reconstruction is an open research problem. Today it is handled mainly by simple statically defined rules, which do not reflect a complexity of the task. Most of these methods do not consider multitasking, which is performed by users on the Web quite often. As a result, relating actions could be reconstructed into different sessions if they were interrupted by another task [79].

The process of session modelling includes the selection of effective representation and persistence approach that should consider further session application (e.g., user behaviour analysis or prediction of future actions). As the user actions acquired by reactive methods without unique user identifiers may originate from different users, methods processing this data should be aware of some uncertainty. For this reason, often additional information such as a site topology [84] or thematic similarity of consecutive actions are used to group user actions clearly [68].

3.2.1 Session Reconstruction

The session reconstruction represents a process of grouping related user actions and joining them into the sessions. For the purposes of user behaviour analysis, the sessions offer more information than individual actions (e.g., identification of typical browsing sequences [142], short and long-term preferences [49], actual user intent [19], behavioural patterns [43]).

Various heuristics, as a maximum time spent in a session, site topology compliance or semantic content of pages are used to reconstruct sessions from user actions [111]. The state-of-the-art approaches often use simple statically defined rules to join the actions into the sessions. The sequence of actions is joined when the time gap between consequent actions is below defined threshold. The actual length of this gap depends on a domain and the specific site [97].

The time approach is insufficient in some situations, because it ignores the relations between the actions [130]. Decision based on time information only could join unrelated actions or separate a session with long page visits. Moreover, if the user spends more time on some page and the next visit occurs too late, these actions are not joined into the same session even if they are related. The active time spent on page may help to reduce this problem [131].

Multiple heuristics of session reconstruction were proposed, classified by the level of user privacy protection:

- *Rule based heuristics* – Simple approach based on the important user action characteristics such as the IP address, the browser fingerprint and visit timestamp. Users are identified by the combination of their characteristics (IP address, browser fingerprint etc.). Actions performed by one user are sorted simply by their access time and divided into sessions by defined rules [144].
- *Temporal heuristics* – The most popular method based on a simple assumption that a session ends if the user does not visit the new page within some defined time interval. It was firstly used in [26] with the threshold set to 25.5 minutes. The authors found out that average gap between two user requests is 9.5 minutes and they extended it by 1.5 of standard deviation. This would, however, make sense only with the normal distribution, which causes that there is only a minimal difference when 20 or 40 minutes gap is used [59].

  As this approach is easy to use, it is very popular and exists in variations from 5 minute [45] up to 30 minutes gaps [87, 113]. Sometimes the personalised size of session gap was used [104]. According to Velasquez and Palade [131], there is no reasonable explanation for usage of popular 30 minutes gap size. An explanation is that the user preferences do not generally change within 20 or 30 minutes thanks to the similar context and user intent. Identification of a user is similarly to rule based heuristics, made by combination of IP address and browser fingerprint. The time oriented heuristics are not able to detect extremely short sessions and similarly the long periods where users work on a same task without browsing or searching on the Web.

- *Topology oriented heuristics* – Users are identified similarly to previous two approaches, by the combination of IP address and browser fingerprint. Page views of a user are then joined into the same session only if there exists a hyperlink connection between pairs consecutive in time. Otherwise, the new session is created [36]. On the one side, this helps to differentiate two users with the same IP address and browser fingerprint (if they browse various parts of the site in the same time). On the other side, the approach fails when several users browse the same part of the site or they cross their paths [40]. Another complication occurs when sites contain narrow parts (sections with only few ways how to get from), e.g., section in news site is accessible from the menu only. In this case, the problem is that many users browse through these bottleneck and it is not possible to differentiate their browsing paths [84].
- *Ontology based heuristics* – The approach assumes that the user always acts with some purpose, which can be mostly identified from the sequence of visited pages. The enrichment of server logs is assumed by ontological information about visited pages. Ontological descriptions are pre-calculated for every page (URL) and also for the pair of pages (hyperlink connection) [71]. Sessions are reconstructed by finding the nearest URLs according to the semantic distance [75]. The disadvantage of this approach is that in a situation when the user clicks on a different page, the approach always creates a new session.
- *Dynamic environment heuristics* – Previous approaches assumed a static content of pages, which is not updated over the time. It is beneficial to take into account the page content updates [106]. Nasraoui et al. asserted the dynamic URL as a valid representation, which is suitable for the website sessions modelling. The comparison between different versions of the same pages is possible only based on their semantic similarity. In this way, the repeated visits of the same page are considered as visits of different pages.
- *Lexical distance heuristics* – The approach is mainly used for tasks where sessions are created from user's search actions. The approach considers a lexical distance of search queries. Its idea is to compare the content of two queries in order to detect changes in the intent [68]. A disadvantage of this approach is the production of high amount of false positive decisions (the split of the actions, which should be joined). This is caused by the fact that users often use completely different queries to search for similar topics [79].

- *Hybrid heuristics* – Hybrid methods are based on the combination of multiple approaches (e.g., temporal and lexical) [51]. Gayo-Avello's approach compared the pairs of consequent queries to border situations (e.g., parallel and different visits and similar visits made with long mutual gap). The combination of previous approaches reduces their problems, but it also brings higher computational and time requirements.

The session reconstruction is trivial for the majority of actions. In such situation all of described approaches perform well. As the session reconstruction is a ambiguous task, it is quite hard to pick a clear winner for the specific sessions (long, interrupted, mixed). As a result, the decision has to be done based on the deep analysis of the specific domain and users characteristics.

3.2.2 Session Representation

Sessions are useful to aggregate information for single actions. To be able to proceed in the user behaviour modelling and to apply various approaches, we need to represent them. There exist three basic session representations [114]:

- *Per page weight representation* – sessions are represented as the weighted vectors $V = [w_1, w_2, \ldots, w_n]$, which weights are based on the time spent on pages (each page is a vector element). These time intervals are normalised [98]. Weighted representation was originally designed as a vector for the similar users search.
- *Graph representation* – session is represented as a sequence of visited pages. These are stored as a graph constructed using a similarity measure and sequence alignment algorithm [57]. Such representation is effective for session storage and search for similar sessions.
- *Representation considering text weighting* – two information types are considered. Primarily, pages are described by their semantic relations based on site partial elements. Based on the time in which user visited the page, the interest for its elements is computed. The interest estimations are stored in the user model as the importance vector [131].

As can be seen, several approaches were proposed to be able to store the sessions in an effective form. These approaches differ mostly in complexity of information they consider and thus also in representation storage size and processing time. The "per page weight representation" is fast and memory efficient, it however considers only an information about the time spent on pages. More sophisticated approaches consider also visits order and semantics.

3.3 Session Intent Identification

User behaviour on the Web subjects to some intent. He/she may want to find some information, read the news or buy some goods. This information is somehow stored in the session, while we can search for mutual actions characteristics. Understanding of these traits and the intent itself is helpful for the fulfilling of user information needs [19].

The session intent identification was firstly researched in [19]. Author focused on the classification of user's informational needs. He described three types of user's needs on the Web – navigational, informational and transactional. This mechanism classifies the sessions on more abstract level than its actual intent, however, it offers a valuable information. The three types identified by Broder are:

- *The navigational session* – is performed, when the user knows the particular site he/she is looking for and wants to directly access it (based on the previous experience).
- *The informational session* – occurs when the user wants to acquire some information from one or multiple sources, e.g., looking for specific information, general looking for actual news or friends' activities on social media or blogs.
- *The transactional session* – represents situations, when the user makes some kind of transaction, e.g., buying some items, downloading a file.

Clearly sessions without a search belong to informational (browsing) or transactional (buying) sessions type [90].

When comparing search and nonsearch sessions, researchers report that 46.7% of actions belongs to the user transactions (e.g., mailing), 19.9% to the Web browsing (e.g., news reading), 16.3% to facts searching (e.g., looking for relevant statements), 13.5% to information gathering (e.g., looking for bus departures) and 1.7% to other non classified actions [74].

In addition to session intent, authors in [130] focused on a topic detection. They claimed that actual user goals depend on many context factors as the user mood, weather, previously visited pages, seen topics or items. Authors proposed an approach for the user intent identification from repeating actions within a session. Their approach uses implicit user feedback, in the form of page visits in already visited items. Authors proposed the factorised Markovian decision process (fMDP), which models every attribute independently. This results to the identification of multiple attributes (Figure 3).

Another approach for intent identification includes topics seen in previous sessions. Often the forgetting mechanism is introduced, e.g., the Latent

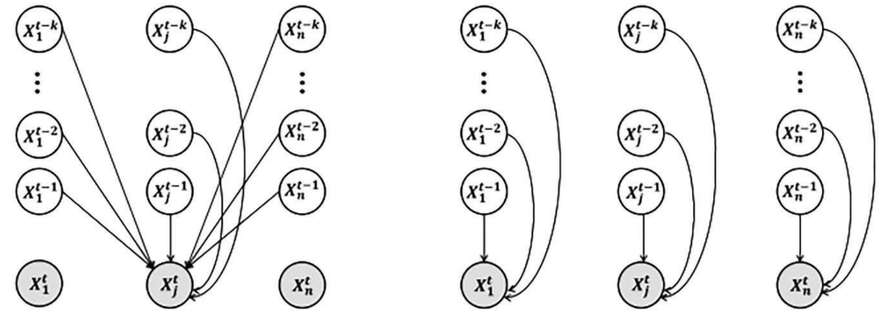


Figure 3 Factorised MDP. Left: standard joint transactional model, which leads to an infeasible model. Right: factorised MDP variant with independent attributes [130].

Dirichlent Allocation model [10] or the Markovian decision process [130]. Wang et al. discussed two forgetting mechanism types – a sliding time window, which includes only visits from defined time interval, and the imitation of the natural human forgetting mechanism [133].

Additionally to various intent of the sessions, there also vary the users who perform them. According to White and Drucker [135], there are several user types based on their cognitive style:

- *Navigators* – 17% of users, who search the information sequentially while they revisit websites frequently with consistent sequence patterns. They have the highest influence on the amount of data collected regardless of their absolute multiplicity.
- *Explorers* – 3% of users. Explorers prefer highly variable behavioural patterns. They often visit new pages with the tendency to navigate strictly through search queries.
- *Other* – 80% of user characteristics by the large variety of behavioural patterns, while every one of them forms only a small percentage.

After identifying of the session intent, it can be used for improving user experience on the site. If we are able to recognise user's aim, we can help him/her for example to find the information faster by prediction of his/her future intent [32], buy the goods he/she prefers, find interesting news, improve search result [126] but also predict user's intent to end the session [73].

To sum it up, the website usage mining represents the most important part of the website mining process from the view of the user behaviour modelling process. The reason is that it allows the most direct description of user behaviour, identification of related actions and also actual user's intent within the site. Together with visited content and site structure, these information represent an input for user modelling process.

4 User Behaviour Modelling

Model generally represents a simplification of the reality. In the context of behaviour modelling, the user model represents user behavioural features, which describe the user preferences, his/her typical actions or differences to other users. Based on its usage, the user model should contain information about [79]:

- User preferences, interests, goals or attitudes
- Proficiencies, knowledge
- Interaction history (user's interaction with the system, performed tasks)
- Stereotypes (e.g., predefined categories)

In this section, we discuss principles of user behaviour modelling. We describe existing model representations and various levels of modelling – the long- as well as the short-term. Finally, we analyse one of the typical problems – the model initialisation.

The task of user modelling is typically studied in the context of the personalization [99], fraud detection [61] or shopping basket prediction [143]. According to Brusilovsky, the user model is defined as a set of information connected with the user behaviour, attitudes and stereotypes [21]. Brusilovsky differentiates two types of Web based user models:

- *Stereotype user model* maps users into one of the predefined groups
- *Overlay user model* reflects user characteristics by adding a layer containing information related to the user and to the domain model for each user individually

Senot et al. represent the user model as a set of pairs $[category, relevance]$, where $relevance \in \langle 0, 1 \rangle$ and describes the user interest for a category. It is calculated based on a time spent with items of specific category or as an average rating of items of category [120].

He described the model precisely by the three attribute types – *Quantity of Affiliation*, *Quantity of Consumption* and *Quantity of Interest*. *Quantity of Affiliation* – QoA represents the level of item affiliation to some semantic concept or category (e.g., article "Men walk on moon", $QoA = \{science = 0.9,\ engineering = 0.7\}$). *Quantity of Consumption* – QoC represents the measure of item consumption by the user, which means an amount of the item that user consumed (e.g., article "Men walk on moon",

$QoC = \{415/524\ words\ read,\ 5/5\ rating\}$). *Quantity of Interest – QoI* is a measure of overall user interest by some semantic concept (e.g., user 1928, $QoI = \{hockey = 0.45,\ science = 0.92\}$). In this meaning user model is constructed as the join of $QoCs$ for the set of items the user interacted with (e.g., rated, read, seen) [120].

Traditional modelling approaches capture mostly the long-term user preferences which describe his/her stable characteristics. This traditional form was later supplemented by modelling of actual user aim or context [77, 115].

The process of user modelling described Barla, who divided it into two parts in his work [11]:

- The user model initialisation (new user)
- The user model maintaining and update (known user)

Brusilovsky [21] described the process of model initialisation and divided it into several steps:

1. Data collection from various sources by the implicit or explicit feedback.
2. The user model inference containing the phase of processing logged user data into higher level (e.g., preferences, interests).
3. Usage of the model (e.g., adaptation or personalization of the content to the user).

In the following subsection, we focus on the second step, which includes the creation of the user model and modelling process itself. At first, we discuss various model representations, which allow us to understand the pros and cons of typically used approaches. Then we focus on the initialisation of user model, which is the key model part. Finally, to follow up actual trends on the Web, we also emphasise the modelling within dynamic domains, where content often changes as we mentioned above.

4.1 User Model Representation

User behaviour can be captured in many ways. There exist implicit models, where preferences are reconstructed from the user sessions in the moment when they are needed (e.g., matrix of ratings). Next, there are explicit models, where preferences are directly stored in the model, which is created and maintained based on the actions. In this way it is possible to effectively use the model for various application tasks (e.g., personalised recommendation, behaviour prediction). The representation of the user model should always consider its further usage. A simple model will not be able to record all required information, too complex model, will require an expensive maintenance, while its potential will be not fully used. According to the literature, most widespread used model representations are:

- *Bag of words* – the oldest and simplest representation [58]. Model is composed of unstructured set of words or terms. The simplicity of the model allows performing operations that are fast (item addition or removal, models union or intersection). On the contrary it does not offer any weights, all items are considered with the equal importance.
- *Vector* – An improvement of the previous approach in the meaning of extending the model by the items importance (weighting). The vector model is similar to general description of the user model presented by Senot et al. [120]. It allows to model the user preferences importance (e.g., words, terms, categories). It was firstly described in [38] as a set of items with binary or integer weights. Similarly, to the bag of words representation, this approach allows simple similarity computation between two models. This process can be optimised by the usage of min hashing or similar comparison techniques [20]. The model maintenance stands in a simple and cheap update of weights. As this model is able to store only attributes of one type, parallel usage of multiple independent vectors is often used for every attribute type. Debnath et al. [39] used this idea to model multiple item characteristics (e.g., movies, actors, directors, genres etc.) and to concurrent model of various user moods.
- *Graph (network based)* – The next level of the user modelling is represented by a graph or a network based approach. It introduces the modelling of relations between items (e.g., hyperlink connection between documents, news articles, a connection between a movie and its director). All items are stored in a single graph. The relations between users and items or similarities between the items itself are expressed by the graph edges (which can be weighted). The model was used to improve the search for similar elements or users [140]. Its disadvantage is the complex structure. The vertexes and operations of modelled graph occur in multiple types and thus the models of comparison process are complex.
- *Ontology based* – The top conceptual level of modelling information about the user. The ontology based model represents an extension of the graph model, as it contains also semantic relations between the model items. In this way, the connections are represented by structures carrying semantical information, which allow a description of advanced relations. Despite the power of ontologies and such a model, it is very expensive to build and maintain ontology based model, and thus it is rarely used [121].

Described user models differ in their express power and computational requirements. As a rule, it is used as simple representation as possible according to the task. Simple approaches typically require less storage space and computational power for building and maintenance. The advantage of complex models is, logically, their high expressive power enabling model usage for advanced tasks.

Nowadays, we are facing a trend of cross system/domain models, where information are shared between multiple sources or sites for improving the user model quality [24, 123]. From the representation point of view, this usually creates sets of independent models or one primary model enriched by information gained from secondary model [30]. In this case, only primary model is used for application purposes.

4.2 User Model Initialisation

Similarly, to other tasks, user modelling also suffers from so-called cold start. in other words, it takes some time to collect enough user actions to model his/her preferences and to be able to further use his/her user model (e.g., for recommendation or prediction). Despite of being researched for many years, it represents a serious long-standing open research problem [53]. When a new user visits the site, he/she expects high quality services and he/she is prone to leave and never return back again if he/she does not get them on actual site.

In previous sections we described the process of usage and website data acquisition. However, for a new user, there are not available any information about his/her previous actions within the actual website. For this reason, the external information could be used for the user model initialisation. The data sources for cross system models [53] are divided into demographic data describing user characteristics (e.g., age, education, sex, salary) [86] and the data from outside user models that capture user behaviour in other sites [89]. The implicit data from external information sources are, however, difficult to acquire and model. This is true especially for reactively identified users without unique user identifiers (it is difficult to clearly join actions between various data sources and no clear identifier to join). In the case of proactively acquired actions, users have different identifiers on different sites, so mapping information from both sources is quite challenging.

The second approach is based on acquiring explicit information directly from the user. The user is typically, asked to rate some items [141] or select preferred items from presented items pairs [56]. If the extensive information is required from the user before offering him/her site services (e.g., recommendation), the chance that user leaves the website increases. To optimise amount of asked questions during initial information retrieval, Sun et al. [128] constructed the question based decision tree. Each tree node contains linear classifier considering previous user answers and selects the next question to ask. Information obtained directly from the user is a valuable source. Its disadvantage is however the necessity to for direct asking and bothering the user. As users are often not willing to offer explicit answers, this source of information is quite uncertain.

The bottleneck of this idea is the user. It is crucial to explain to the user that offering information about him/her (or allowing usage of additional information source) will increase his/her own user experience on the website, the user model is used in.

4.3 Dynamic Domain Modelling

Modelling of various user interests, by reacting to frequent changes and maintaining model information up to date, is the most challenging task in dynamic domains (e.g., news [133]). This, however, opens questions such as how to deal with dynamically changing user interests, or how to model interests weakening or even forgetting.

4.3.1 Multiple Layer Modelling

In dynamic domains, the pages are often updated, they are actual only for a while and change quickly. That's the reason why user preferences are modelled on multiple levels. The two level model variant is preferred by multiple authors in [14, 33, 133, 136, 139], who identically divide model to *long-* and *short-term* parts. Another approach proposed in [145] added third level of *medium-term* preferences.

Wang et al. used memory based user model. This approach is used to model preferences of individual topics, which dynamically reacts to changes in user interests [133]. Next, these authors proposed extension of the model by including multiple parameters as the absorbing and forgetting, timescale and learning strength (used to simulate human characteristics). Modelling user interest for topics is typically realised by methods considering a time window [63, 133] or a forgetting curve [31, 88]. These approaches also used a time spent on page, measure of scrolling, clicking, bookmarking, printing or selecting text [105].

The importance of preference forgetting is pointed also in [145]. Zhou et al. used the ZGrapher algorithm to analyse memory characteristics and a Forgetting and re-energising user preference algorithm to model user preferences, its changes and forgetting over the time. This model is intended to imitate the human mind. If an item that has been placed in short-term layer, will be not visited soon again, it will be forgotten. The medium-term layer keeps items in memory longer. The most stable are logically the long-term interests.

Billsus and Pazzani [14] proposed a user model for news articles recommendation. Their model is based on a combination of distance based methods as k-nearest neighbours [37] and Bayesian methods [46]. In this way, the model captures user's both short-term and long-term interests and switch between them [33]. Short-term part is focused on the track ongoing events, which might be highly interesting for the user. The long-term model part tracks topics, which do not need to be actually trendy, but the user has a long time interest in them.

With the idea of two level user model agrees also Xiang et al., who say that for the recommendation it is important to consider the time aspect as well (not only similar users). For this reason, these authors distinguish if users chose items via long or short-term preferences [139]. Mourao et al. found out that past long-term relevance is more promising information for identifying unexpected items. On the contrary, the short-term behaviour allows to identify high percentages of actually consumed and relevant items [102]. Some authors use time as a universal dimension, which may be compared between all users [127]. Xiang et al. however, argue that time has only a local impact and should not be compared. To deal with these problems [139] proposed a simple graph based user model – Session based temporal graph, which can capture long-term as well as short-term preferences (Figure 4). We believe that it is beneficial to distinguish between long- and short-term behaviour but their comparison could bring another information.

In addition to user personal preferences, dynamically changing domains are also influenced by actual trends, i.e., topics which are preferred by high number of users for a limited time. The influence of global trends to personal user interests on social network (Twitter) was researched in [49]. Authors in their study reported that user personal interests have high importance for the recommendation task. These interests cannot be replaced by tracking of public global trends. The trends can be however used as an additional source of input information for a modelling process [49].

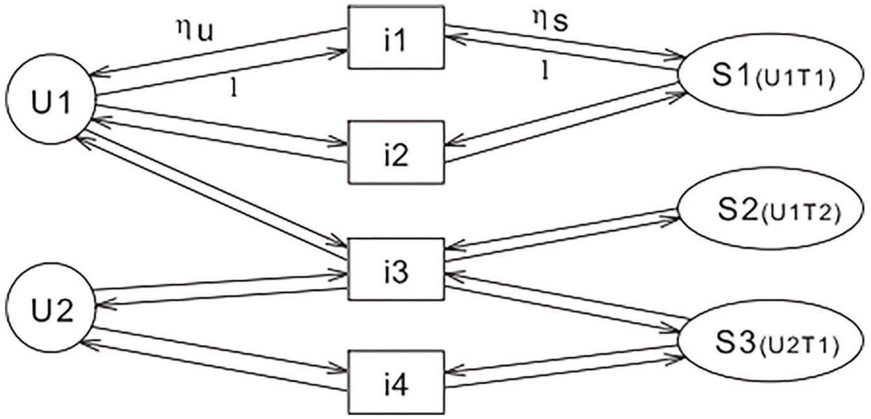


Figure 4 Visualization of simple STG graph showing 2 users who interact with 3 (2) items in 2 (1) session. In the graph, the set of vertexes $U = \{U_1, U_2, \cdots\}$ represents users, the set of vertexes $i = \{i_1, i_2, \cdots\}$ represents items and set of vertexes $S = \{S_1, S_2, \cdots\}$ represents sessions. The edges (user to item and item to session) are oriented. Sessions $S_x = [U_x, T_y]$ store the information about its author (user U_x) and the time of interaction T_y. In this representation, user vertexes represent long-term preferences, session nodes represent short-term preferences. The model allows adjusting the window size used for session selection [139].

The long- and the short-term behaviour is, however, always related and there is impossible to set the strict threshold to divide them. On the contrary to previous approaches distinguishing between long- and short-term behaviour, in [143] authors use the long-term user purchase history dynamically updated by short-term changes to identify the latent shopping basket preferences. They found out that it is insufficient to focus on the latest purchases only, but the next basket prediction requires latent connection of both long- and short-term behavioural types.

5 Conclusions

The user behaviour can be described by multiple patterns and statistical distributions [29]. Despite this characteristic, it is still a complex set of actions where every decision is a result of the subconscious consideration of multiple factors (e.g., actual information need, preferences or previous knowledge). In addition to these factors which depend mainly on the user himself, also external factors influence the behaviour as the actual context, website structure and a frequency of its changes. To be able to successfully describe the user behaviour, these factors have to be considered and their influence evaluated.

The user modelling depends on the quality and amount of available data. The Web mining process typically uses generally available sources (website itself and standard transaction server logs). This allows extracting information about the website structure, content of its pages and also about the site usage. User actions logically represent the most direct way of user behaviour description, but the other sources (e.g., actual trends, events, stimuli from other users) help us to understand the context of the user. The disadvantage of the contextual sources is however that they are relatively specific and thus could be unavailable for some websites, which decreases the model re-usability.

The result of the mining process is the set of pre-processed user actions. Individual actions joined into the sessions, offer additional information about the user intent, his/her path across the site and gained knowledge. There exist multiple heuristics to identify related actions, but the most widespread is the time based, which combines simplicity with sufficient quality.

In recent years, the process of user modelling has been adapted to changes of the Web. Nowadays, websites are dynamically changing, pages are created, often updated and they remain actual only for a while. An example are the social media and news websites, where the content is updated as a continuous data stream. For this reason, the user modelling does not rely only on the traditional long-term modelling but it is often combined with the short-term one. The short-term models emphasise the freshness of the information over its stability.

The advantage of long-term behaviour modelling is the ability to capture stable user preferences. Typical application is the personalised recommendation. The disadvantage is the long reaction time to user change of interests and its inability to react to the actual user behaviour.

The opposite occurs in the short-term modelling. The specific user action is influenced also by actual context, needs or trends (in addition to his/her stable preferences and knowledge). For this reason, it is quite challenging to capture and model recent user behaviour. The innovative ideas are required to handle all noisy short-term factors and extract the actual user preferences. Short-term modelling itself however cannot describe stable user preferences and thus should be combined with the long-term approach.

A combination of both approaches brings long-term stability and short-term actuality and dynamicity that allows us to model a complex user behaviour in online time. In this way a fast analysis of previous user actions hand by hand with long-term preferences is possible. As a result, such model improves the user experience on the website, his/her satisfaction and potentially increases website revenue.

The importance of the short-term modelling is also supported by the fact that majority of website users is anonymous or occasional (e.g., for news, e-shops, business sites). For these users, it is impossible to build well-performing long-term user models. Their models can be initialised only by demographic characteristics or some predefined stereotypes. On the contrary, the short-term model can be built from a low number of actions in the online time, within the active session.

The user models considering the latest user behaviour allows us to react to its changes, pages updates and actual user context in online time. As these models offer up to date information, they can be used for the immediate adaptation, personalization or generally for tasks increasing the actual user experience on the website and his/her satisfaction. These allows us to design Web application reacting on the actual user behaviour.

We believe that hand by hand with the widespreading of special hardware (e.g. eye-trackers) to end-users, there will be more information available about users and their short-term context. In this way, we will be able to describe actual user situation in qualitatively better way, which will lead to more precise short-term models and thus to the improvement of Web services.

The future of the user modelling on the Web seems to be quite challenging. On the one hand, we expect tremendous increase of information related to the user context. We can used increasing share of mobile devices and their sensors [12]. Moreover, the number of wearable devices in 2020 is expecting to exceed 830 million[2]. On the other hand, user privacy concerns will definitely play a crucial role in the design and application of user-related information and methods.

Acknowledgements

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0508 and APVV-17-0267; the Scientific Grant Agency of the Slovak Republic, grant No. VG 1/0646/15 and VG 1/0667/18; the Research and Development Operational Programme for the project ITMS 26240120039, co-funded by the European Regional Development Fund.

[2] https://www.statista.com/topics/1556/wearable-technology/

References

[1] Abel, F., Herder, E., Houben, G. J., Henze, N., and Krause, D. (2013). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2–3), 169–209.

[2] Adar, E., Teevan, J., Dumais, S. T., and Elsas, J. L. (2009). The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 282–291). ACM.

[3] Alexander, J., and Cockburn, A. (2008) An empirical characterisation of electronic document navigation. In *Proceedings of Graphics Interface 2008 (GI '08)*, Canadian Information Processing Society, Toronto, Ontario, Canada, (pp 123–130).

[4] AlSumait, L., Barbará, D., and Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Eighth IEEE International Conference on Data Mining. ICDM'08.* (pp. 3–12). IEEE.

[5] Armstrong, S., Church K., and Pierre I. (2014). Natural Language Processing Using Very Large Corpora, Springer-Verlag, Berlin, Heidelberg.

[6] Aswani, R., Ghrera, S. P., Chandra, S., and Kar, A. K. (2017). Outlier Detection Among Influencer Blogs Based on off-Site Web Analytics Data. In *Conference on e-Business, e-Services and e-Society* (pp. 251–260). Springer, Cham.

[7] Baeza-Yates, R., and Boldi, P. (2010). Web structure mining. In *Advanced Techniques in Web Intelligence-I* (pp. 113–142). Springer, Berlin, Heidelberg.

[8] Baeza-Yates, R., Castillo, C., and Efthimiadis, E. (2007), Characterization of national Web domains, *ACM Transactions on Internet Technology,* 7(2), 1–32.

[9] Baeza-Yates, R., and Poblete, B. (2006). Dynamics of the Chilean web structure, *Computer Networks: The International Journal of Computer and Telecommunications Networking – Web dynamics*, 50(10), 1464–1473.

[10] Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., and Bevacqua, A. (2013). Probabilistic topic models for sequence data. *Machine Learning*, 93(1), 5–29.

[11] M. Barla (2010), Towards Social-based User Modeling and Personalization, PhD thesis. Faculty of Informatics and Information Technologies STU, Bratislava, Slovakia.

[12] Barla, M. (2011). Towards social-based user modeling and personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1), 52–60.

[13] Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping Multidimensional Data* (pp. 25–71). Springer, Berlin, Heidelberg.

[14] Billsus, D., and Pazzani, M. J. (2007). Adaptive news access. In *The Adaptive Web* (pp. 550–570). Springer, Berlin, Heidelberg.

[15] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine LEARNING Research*, 3, 993–1022.

[16] Blei, D. M., and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). ACM.

[17] Bondy, J. A., and Murty, U. S. R. (1976). Graph theory with applications (Vol. 290). London: Macmillan.

[18] Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.

[19] Broder, A. (2002). A taxonomy of web search. In *ACM SIGIR Forum*, 36(2), 3–10. ACM.

[20] Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences*. Proceedings (pp. 21–29). IEEE.

[21] Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling AND User-Adapted Interaction*, 6(2–3), 87–129.

[22] Cahn, A., Alfeld, S., Barford, P., and Muthukrishnan, S. (2016). An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web,* (*WWW '16*), (pp. 891–901). International World Wide Web Conferences Steering Committee.

[23] Cai, D., Yu, S., Wen, J. R., and Ma, W. Y. (2003). Vips: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft Research.

[24] Cantador, I., Fernández-Tobías, I., Berkovsky, S., and Cremonesi, P. (2015). Cross-domain recommender systems. In *Recommender Systems Handbook* (pp. 919–959). Springer, Boston, MA.

[25] Carpineto, C., and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR),* 44(1), 1.

[26] Catledge, L. D., and Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web, in *Proceedings of the Third International World-Wide Web conference on Technology, tools and applications*, 27, Elsevier, 1065–1073.

[27] Chakrabarti, S., et al. (1999). Mining the link structure of the World Wide Web. *IEEE Computer*, 32(8), 60–67.

[28] Chang, J., Rosenn, I., Backstrom, L., and Marlow, C. (2010). ePluribus: Ethnicity on social networks, in *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, *10*, 18–25. AAAI Press.

[29] Chovanak, T., Kassak, O., Kompan, M., and Bielikova, M. (2018). Fast Streaming Behavioural Pattern Mining. *New Generation Computing*, 1–27.

[30] Cena, F., Gena, C., and Picardi, C. (2016). An Experimental Study in Cross-Representation Mediation of User Models. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, (WWW '16).* (pp. 289–290). ACM.

[31] Cheng, Y., Qiu, G., Bu, J., Liu, K., Han, Y., Wang, C., and Chen, C. (2008). Model bloggers' interests based on forgetting mechanism. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08).* (pp. 1129–1130). ACM.

[32] Cheng, Z., Gao, B., and Liu, T. Y. (2010). Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10).* (pp. 221–230). ACM.

[33] Chiu, B. C., and Webb, G. I. (2005), Using decision trees for agent modeling: improving prediction performance, *User Modeling and User-Adapted Interaction*, 8, Springer, 131–152.

[34] Cho, J., and Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB '00)*, Morgan Kaufmann Publishers Inc., 200–209.

[35] Chudá, D., and Krátky, P. (2014). Usage of computer mouse characteristics for identification in web browsing. In *Proceedings of the 15th International Conference on Computer Systems and Technologies (CompSysTech '14),* ACM, 218–225.

[36] Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 5–32.

[37] Cover, T., and Hart, P. (1967). Nearest Neighbor pattern classification, *IEEE Transactions on Information Theory*, 13, 21–27.

[38] De Bra, P., et al. (2003). AHA! The adaptive hypermedia architecture. In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia (HYPERTEXT '03)*, ACM, 81–84.

[39] Debnath, S., Ganguly, N., and Mitra, P. (2011). Feature weighting in content based recommendation system using social network analysis, in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, ACM, 1041–1042.

[40] Dell, R. F., Roman, P. E., and Velasquez, J. D. (2008). Web user session reconstruction using integer programming. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 01 (pp. 385–388). IEEE Computer Society.

[41] Desikan, P., and Srivastava, J. (2004). Mining temporally evolving graphs. In *Proceedings of the 6th WEBKDD Workshop in Conjunction with the 10th ACM SIGKDD Conference, (WebKDD'04)*. (Vol. 22).

[42] Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar D., and Tomkins, A. (2016). Self-similarity in the web, *ACM Transactions on Internet Technology (TOIT)*, 2(3), 205–223.

[43] Doddegowda, B. J., Raju, G. T., and Manvi, S. K. S. (2016). Extraction of behavioral patterns from pre-processed web usage data for web personalization, in *IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT)*, 494–498.

[44] Doudalis, S., Mehrotra, S., Haney, S., and Machanavajjhala, A. (2016). Releasing True Data with Formal Privacy Guarantees. In *Privacy-Preserving IR Workshop at SIGIR.*.

[45] Downey, D., Dumais, S. T., and Horvitz, E. (2007). Models of searching and browsing: languages, studies and applications, in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, Morgan Kaufmann Publishers Inc., 2740–2747.

[46] Duda, R., and Hart, P. (1973). Pattern Classification and Scene Analysis, Wiley and Sons.

[47] Ferrara, E., De Meo, P., Fiumara, G., and Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-based Systems*, 70, 301–323.

[48] Fetterly, D., Manasse, M., Najork, M., and Wiener, J. (2003). A large-scale study of the evolution of web pages. In *Proceedings of the 12th*

International Conference on World Wide Web (WWW '03), (pp. 669–678). ACM.

[49] Gao, Q., Abel, F., Houben, G. J., and Tao, K. (2011). Interweaving trend and user modeling for personalized news recommendation. In *2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, (Vol. 1, pp. 100–103). IEEE.

[50] Gambhir, M., and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1), 1–66.

[51] Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12), 1822–1843.

[52] Goel, S., Hofman, J. M., and Sirer, M. I. (2012). Who Does What on the Web: A Large-Scale Study of Browsing Behavior, in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, AAAI*, 1–8.

[53] Gope, J., and Jain, S. K. (2017). A survey on solving cold start problem in recommender systems, in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 133–138.

[54] Granka, L., Feusner, M., and Lorigo, L. (2008). Eye monitoring in online search. In *Passive eye monitoring* (pp. 347–372). Springer, Berlin, Heidelberg.

[55] Grannis, K., and Davis, E. (2009). China internet network information center, in *14th statistical survey report on the internet development of china 2009*. According to http://www.cnnic.net.cn/uploadfiles/pdf/2009/10/13/94556.pdf

[56] Graus, M. P., and Willemsen, M. C. (2015). Improving the User Experience during Cold Start through Choice-Based Preference Elicitation, in *Proceedings of the 9th ACM Conference on Recommender Systems – RecSys '15*, ACM, 273–276.

[57] Gündüz, Ş., and Özsu, M. T. (2003). A web page prediction model based on clickstream tree representation of user behavior, *KDD '03: Proceedings of the 9th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 535–540.

[58] Harris, Z. S. (1954). Distributional Structure, Word, 10(2/3), 146–162.

[59] Herder, E. (2007). An Analysis of User Behavior on the Web – Understanding the Web and its Users VDM Verlag.

[60] Hill, F., Cho, K., and Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data, In *Proceedings of NAACL-HLT*, 1367–1377.

[61] Hilas, C. S., and Sahalos, J. N. (2006)., Testing the Fraud Detection Ability of Different User Profiles by Means of FF-NN Classifiers, in *Proceedings of the 16th International Conference on Artificial Neural Networks*, Lecture Notes in Computer Science, Part II, 4132, Springer, 872–883.

[62] Hu, J., Zeng, H. J., Li, H., Niu, C., and Chen, Z. (2007, Demographic prediction based on user's browsing behavior, in *Proceedings of the 16th International Conference on World Wide Web (WWW)*, ACM, 151–160.

[63] Huang, X., Yang, Y., Hu, Y., Shen, F., and Shao, J. (2016). Dynamic User Attribute Discovery on Social Media, In *Web Technologies and Applications: 18th Asia-Pacific Web Conference APWeb*, Springer, 256–267.

[64] Huberman, B., Pirolli, P., Pitkow, J., and Lukose R. M. (1998). Strong regularities in world wide web surfin, 280(5360), Science, 95–97.

[65] Huberman B. A., and Wu, F. (2007), The economics of attention: maximizing user value in information-rich environments, *ADKDD '07: Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, ACM, 16–20.

[66] Huntington, P. N., and Jamali, H. R. (2008). Website usage metrics: A reassessment of session data, *Information Processing and Management: an International Journal*, 44(1), 358–372.

[67] Ipeirotis, P. G., and Gravano, L. (2004). When one sample is not enough: improving text database selection using shrinkage, *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, ACM, 767–778.

[68] Jansen, B. J., Spink, A., Blakely, C., and Koshman, S. (2007). Defining a session on web search engines: Research articles, *Journal of the American Society for Information Science and Technology*, 58, 862–871.

[69] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search, *ACM Transactions on Information Systems (TOIS)*, 25(2), ACM, 7.

[70] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.

[71] Jung, J. J., and Jo, G. S. (2004). Semantic outlier analysis for sessionizing web logs, ECML PKDD 2004 – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, 13–25.

[72] Kang, J., and Lee, H. (2017). Modeling user interest in social media using news media and wikipedia. *Information Systems*, 65, 52–64.

[73] Kassak, O., Kompan, M., and Bielikova, M. (2016). Student behavior in a web-based educational system: Exit intent prediction. *Engineering Applications of Artificial Intelligence*, 51, 136–149.

[74] Kellar, M., Watters, C., and Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7), 999–1018.

[75] Khasawneh, N., and Chan, C. (2006). Active user-based and ontology-based web log data preprocessing for web usage mining, IEEE / WIC / ACM International Conference on Web Intelligence (WI 2006), IEEE, 325–328.

[76] Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, ACM, 668–677.

[77] Kompan, M., and Bieliková, M. (2013). Context-based Satisfaction Modelling for Personalized Recommendations, In *Proceedings of the 8th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2013)*, IEEE, 33–38.

[78] Kosala, R., and Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1), 1–15.

[79] Kramar, T., Barla, M., and Bieliková, M. (2013). Personalizing Search Using Socially Enhanced Interest Model Built from the Stream of User's Activity. *J. Web Eng.,* 12(1&2), 65–92.

[80] Krátky, P., and Chudá, D. (2018). Recognition of web users with the aid of biometric user model. *Journal of Intelligent Information Systems*, 1–26.

[81] Kumar, R., and Tomkins, A. (2010). A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World wide web* (pp. 561–570). ACM.

[82] Labaj, M., and Bieliková, M. (2013). Tabbed browsing behavior as a source for user modeling. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 388–391). Springer, Berlin, Heidelberg.

[83] Lee, Y., Rajasekar, V. C. S., and Kasula, P. R. (2018). Accessibility of Website for Visually Challenged: Combined Tree Structure and XML Metadata, *GSTF Journal on Computing (JoC),* 2(1), 1–11.

[84] Li, Y., Feng, B., and Mao, Q. (2008). Research on path completion technique in web usage mining. In *International Symposium on Computer Science and Computational Technology*, *ISCSCT'08*. (Vol. 1, pp. 554–559). IEEE.

[85] Li, X., Ouyang, J., Lu, Y., Zhou, X., and Tian, T. (2015). Group topic model: organizing topics into groups. *Information Retrieval Journal*, 18(1), 1–25.

[86] Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4), 2065–2073.

[87] Liu, C., White, R. W., and Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and development in Information Retrieval* (pp. 379–386). ACM.

[88] Liu, K., Chen, W., Bu, J., Chen, C., and Zhang, L. (2007). User Modeling for Recommendation in Blogspace, *WI-IATW '07 Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops*, IEEE, 79–82.

[89] Loh, S., Lorenzi, F., Granada, R., Lichtnow, D., Wives, L. K., and de Oliveira, J. P. M. (2009). Identifying Similar Users by their Scientific Publications to Reduce Cold Start in Recommender Systems, in *Proceedings of the 5th International Conference on Web Information Systems and Technologies – WEBIST*, 593–600.

[90] Loyola, P., Liu, C., and Hirate, Y. (2017). Modeling User Session and Intent with an Attention-based Encoder-Decoder Architecture. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 147–151). ACM.

[91] Lupu, R. G., and Ungureanu, F. (2013). A survey of eye tracking methods and applications. *Buletinul Institutului Politehnic din Iasi, Automatic Control and Computer Science Section*, 3, 72–86.

[92] Marujo, L., Ling, W., Ribeiro, R., Gershman, A., Carbonell, J., de Matos, D. M., and Neto, J. P. (2016). Exploring events and distributed representations of text in multi-document summarization. *Knowledge-Based Systems*, 94, 33–42.

[93] Manzato, M. G., et al.,. (2016). Mining unstructured content for recommender systems: an ensemble approach. *Information Retrieval Journal*, 19(4), 378–415.

[94] Mathew, L., Elias, A., and Ravi, C. (2016). Total privacy preservation and search quality improvement in personalized web search. *Journal of Web Engineering*, 15(5–6), 465–483.

[95] Mezghani, M., Zayani, C. A., Amous, I., and Gargouri, F. (2012). A user profile modelling using social annotations: a survey. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 969–976). ACM..

[96] Mihalcea, R., and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processi*ng. 404–411.

[97] Mihalkova, L., and Mooney, R. (2009). Learning to disambiguate search queries from short sessions. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 111–127). Springer, Berlin, Heidelberg.

[98] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6(1), 61–82.

[99] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2001, November). Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd International Workshop on Web Information and Data Management* (pp. 9–15). ACM.

[100] Mohan, K., Kurmi, J., and Kumar, S. (2017). A Survey on Web Structure Mining. *International Journal of Advanced Research in Computer Science*, 8(3), 227–232.

[101] Moro, R., and Bielikova, M. (2015). Utilizing Gaze Data in Learning: From Reading Patterns Detection to Personalization,in *Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*, Springer, 1–4.

[102] Mourão, F., Rocha, L., Araújo, C., Meira Jr, W., and Konstan, J. (2017). What surprises does your past have for you?. *Information Systems*, 71, 137–151.

[103] Rosenfeld, L., and Morville, P. (2002). *Information architecture for the world wide web*. “O’Reilly Media, Inc.”.

[104] Murray, G. C., Lin, J., and Chowdhury, A. (2006). Identification of user sessions with hierarchical agglomerative clustering. In *Proceedings of*

the American Society for Information Science and Technology, 43(1), 1–9.

[105] Mushtaq, N., Werner, P., Tolle, K., and Zicari, R. (2004). Building and Evaluating Non-obvious User Profiles for Visitors of Web Sites, in *Proceedings of the IEEE International Conference on E-Commerce Technology (CEC '04)*, IEEE, 9–15.

[106] Nasraoui, O., Soliman, M., Saka, E., Badia, A., and Germain, R. (2008) Aweb usage mining framework for mining evolving user profiles in dynamic web sites, *IEEE Trans. on Knowl. and Data Eng*., 20(2), IEEE, 202–215.

[107] Ntoulas, A., Cho, J., and Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective, in *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*, ACM, pp. 1–12.

[108] Obendorf, H., Weinreich, H., Herder, E., and Mayer, M. (2007). Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 597–606). ACM.

[109] Olston, C., and Pandey, S. (2008). Recrawl scheduling based on information longevity, in *Proceeding of the 17th International Conference on World Wide Web (WWW '08)*, ACM, 437–446.

[110] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

[111] Patel, P., and Parmar, M. (2014). Improve heuristics for user session identification through web server log in web usage mining. *International Journal of Computer Science and Information Technologies*, 5(3), 3562–3565.

[112] Paukkeri, M. S., and Honkela, T. (2010). Likey: Unsupervised language-independent keyphrase extraction, in *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, Association for Computational Linguistics, 162–165.

[113] Radlinski, F., and Joachims, T. (2005). Query chains: learning to rank from implicit feedback. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 239–248). ACM.

[114] Roman, R. P. A. (2011), Web User Behavior Analysis, PhD thesis. Universidad de Chile, Chile.

[115] da Rosa, J. H., Barbosa, J. L., and Ribeiro, G. D. (2016). ORACON: An adaptive model for context prediction. *Expert Systems with Applications*, 45, 56–70.

[116] Ruzza, M., Tiozzo, B., Mantovani, C., D'Este, F., and Ravarotto, L. (2017). Designing the information architecture of a complex website: A strategy based on news content and faceted classification. *International Journal of Information Management*, 37(3), 166–176.

[117] Sadagopan, N. and Li, J. (2008). Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 885–894), ACM.

[118] Schneider-Mizell, C.M. and Sander, L.M. (2009). A generalized voter model on complex networks. Technical Report Department of Physics, University of Michigan.

[119] Scott, S. and Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99),* Morgan Kaufmann Publishers Inc., pp. 379–388.

[120] Senot, C., Kostadinov, D., Bouzid, M., Picault, J., Aghasaryan, A. and Bernier, C. (2010). Analysis of strategies for building group profiles. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 40–51). Springer, Berlin, Heidelberg.

[121] Sieg, A., Mobasher, B. and Burke, R. (2007). Ontological user profiles for representing context in web search. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops,* (pp. 91–94), IEEE.

[122] Simko, J. and Vrba, J. (2018). Screen recording segmentation to scenes for eye-tracking analysis. *Multimedia Tools and Applications*, 1–25.

[123] Simko, M. and Bielikova, M. (2018). Lightweight domain modeling for adaptive web-based educational system. *Journal of Intelligent Information Systems*, 1–26.

[124] Sukumar, P., Robert, L. and Yuvaraj, S. (2016). Review on modern Data Preprocessing techniques in Web usage mining (WUM). In *International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS),* (pp. 64–69), IEEE.

[125] Spiliopoulou, M., Mobasher, B., Berendt, B. and Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, *15*(2), 171–190.

[126] Strohmaier, M., Kröll, M. and Körner, C. (2009). Intentional query suggestion: making user goals more explicit during search. In *Proceedings of the 2009 Workshop on Web Search Click Data* (pp. 68–74), ACM.

[127] Sun, J., Faloutsos, C., Papadimitriou, S. and Yu, P.S. (2007). Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 687–696), ACM.

[128] Sun, M., Li, F., Lee, J., Zhou, K., Lebanon, G. and Zha, H. (2013). Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the 6th ACM international Conference on Web Search and Data Mining* (pp. 445–454), ACM.

[129] Takalikar, V. and Joshi, P. (2016). Inter-page access metrics for web site structure and performance. In *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT),* (pp. 196–203), IEEE.

[130] Tavakol, M. and Brefeld, U. (2014). Factored MDPs for detecting topics of user sessions. In *Proceedings of the 8th ACM Conference on Recommender Systems* (pp. 33–40), ACM.

[131] Velásquez, J.D. and Palade, V. (2008). *Adaptive web sites: A knowledge extraction from web data approach* (Vol. 170), Ios Press.

[132] Wang, P., Qian, Y., Soong, F.K., He, L. and Zhao, H. (2016). Learning distributed word representations for bidirectional lstm recurrent neural network. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 527–533).

[133] Wang, W., Zhao, D., Luo, H. and Wang, X. (2013). Mining user interests in web logs of an online news service based on memory model. In *2013 IEEE 8th International Conference on Networking, Architecture and Storage (NAS),* (pp. 151–155), IEEE.

[134] Wartena, C. and Brussee, R. (2008). Topic detection by clustering keywords. In *19th International Workshop on Database and Expert Systems Application, 2008 (DEXA'08),* (pp. 54–58), IEEE.

[135] White, R.W. and Drucker, S.M. (2007). Investigating behavioral variability in web search. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 21–30), ACM.

[136] Widyantoro, D.H., Ioerger, T.R. and Yen, J. (2001). Learning user interest dynamics with a three−descriptor representation. *Journal of the American Society for Information Science and Technology*, *52*(3), 212–225.

[137] Witten, I.H., Bray, Z., Mahoui, M. and Teahan, W.J. (1999). Text mining: A new frontier for lossless compression. In *Proceedings of the Conference on Data Compression (DCC '99)*, IEEE, 198–207.

[138] Won, S.S., Jin, J. and Hong, J.I. (2009). Contextual web history: using visual and contextual cues to improve web browser history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1457–1466), ACM.

[139] Xiang, L., Yuan, Q., Zhao, S., Chen, L., Zhang, X., Yang, Q. and Sun, J. (2010). Temporal recommendation on graphs via long-and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 723–732), ACM.

[140] Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. (2008). Exploring folksonomy for personalized search. In *Proceedings of the 31st Annual International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 155–162). ACM.

[141] Yang, S. H., Long, B., Smola, A. J., Zha, H. and Zheng, Z. (2011). Collaborative competitive filtering: learning recommender using context of user choice. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11),* ACM, 295–304.

[142] Yang, Y. C. (2010). Web user behavioral profiling for user identification. *Decision Support Systems*, 49(3), 261–271.

[143] Yu, F., Liu, Q., Wu, S., Wang, L. and Tan, T. (2016). A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR ‘16), ACM, 729–732.

[144] Zawodny, J.D. (2002). Linux apache web server administration, Sybex, 2nd Edition.

[145] Zhou, B., Zhang, B., Liu, Y. and Xing, K. (2011). User model evolution algorithm: forgetting and reenergizing user preference. In *2011 IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing* (pp.444–447). IEEE.

Biographies

Ondrej Kassak is currently a postdoc at the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava. He received PhD. degree in 2018 from the same university. His research interests are in the areas of user modelling and personalized recommendation.

Michal Kompan is currently an associate professor at the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava. He received PhD. degree in 2014 from the same university. His research interests are in the areas of personalized recommenders for single or group of users and user modelling. He is a member if IEEE Computer Society and ACM.

Maria Bielikova received her Master degree (with summa cum laude) in 1989 and her PhD. degree in 1995, both from the Slovak University of Technology in Bratislava. Since 2005, she has been a full professor, presently at the Institute of Informatics and Software Engineering, Slovak University of Technology. Her research interests are in the areas of software web-based information systems, especially personalized context-aware web-based systems including user modeling and social networks. She co-authored over 70 papers in international scientific journals and she is editor of more than 50 proceedings. She is a senior member of IEEE Computer Society, ACM and International Society for Web Engineering.