Int. J. Technology Enhanced Learning, Vol. X, No. Y, XXXX

Considerations for recruiting contributions to anonymised data sets

Eelco Herder* and Ricardo Kawase

L3S Research Center,
University of Hannover,
D-30167 Hannover, Germany
Email: herder@L3S.de
Email: kawase@L3S.de
*Corresponding author

Abstract: For the improvement and evaluation of learning management systems as well as concepts and tools for personal history, the availability of suitable data has become increasingly important. In this paper we discuss considerations for creating such a data set and report on our experiences while recruiting contributions for the Web History Repository project (WHR). We focus on measures and techniques for addressing privacy issues and strategies for motivating people to contribute. We briefly report on the outcomes of the first rounds of analysis of the data set and discuss implications for educational data mining.

Keywords: web usage; data set; recruitment; motivation; privacy.

Reference to this paper should be made as follows: Herder, E. and Kawase, R. (XXXX) 'Considerations for recruiting contributions to anonymised data sets', *Int. J. Technology Enhanced Learning*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Eelco Herder is a Senior Researcher in the L3S Research Center, Hannover, Germany. In 2006, he completed his PhD from the University of Twente, the Netherlands on the analysis of user interaction with the world wide web. His current research activities are focused on web usage analysis, user profiling in the social web and the development of tools for Personal Information Management.

Ricardo Kawase is a PhD student in L3S Research Center, Hannover, Germany. He has been actively involved in various projects in the field of technology-enhanced learning. His research interests focus on human-computer interaction, user behaviour, social media and the semantic web.

1 Introduction

E-learning has evolved together with the web. Traditionally, e-learning was regarded as learners interacting with a monolithic Learning Management System (LMS), and all activities took place within the system. Even though dedicated systems such as Moodle[1] and Blackboard[2] are still widely used, these systems are increasingly being complemented by popular Web 2.0 services, such as Facebook, Twitter, Skype and Google Docs[3].

With the growing importance of e-learning, interest in data mining of user activities in LMS has increased (Romero et al., 2008). Similar to the more general field of web usage mining, a bottleneck for educational data mining is the limited availability and generalisability of data sets of learner activities.

Educational data mining and learning analytics traditionally focus on student interactions with course material. However, as these interactions are often interwoven with 'regular' web activities such as consulting search engines and reference material, it may be needed to consider these interactions as well (Butoianu et al., 2010). The combination of educational data mining and web usage mining gains momentum and is often referred to with the term 'learning analytics' (New Media Consortium, 2012).

The creation of data sets – whether educational or 'general' – is a cumbersome task, which is often complicated by privacy issues and difficulties in motivating prospective participants to invest their time and effort. In this paper, we discuss considerations for tackling these issues. The focus of this paper is not on educational data mining per se, but on the more general field of web usage mining – still, as we will see these issues have an impact on educational data mining and learning analytics as well. Further, we report our experiences in building an anonymised data set of web usage data, the Web History Repository (WHR), in order to provide guidelines for researchers who intend to recruit participants for creating data sets.

The remainder of this paper is structured as follows. In Sections 2 and 3 we briefly explain how e-learning and regular web use is interrelated. In Section 4 we discuss past findings on how people visit and revisit web content. In Sections 5 and 6 we provide a short overview on the process of (educational) data mining and the types of user data that are relevant. Considerations on privacy issues and the motivation of participants are discussed in Sections 7 and 8. Finally, we report on our experiences in creating the WHR in Section 9 and conclude the paper with a discussion in Section 10.

2 E-Learning 2.0

'Modern' e-learning takes place in a regular web browser, and is often intertwined with 'regular' web use – varying from using references such as Google and Wikipedia to interacting with friends via Facebook and checking the news on a regular basis (Teevan et al., 2008; Butoianu et al., 2010). In other words, e-learning systems are more and more regarded and used as 'normal web applications'.

For this trend, Downes (2005) introduced the term 'E-learning 2.0', which he conceives as a learner-centred design that is used by digital natives, who syndicate, absorb and share information from multiple sources simultaneously. The term 'E-learning 2.0' has caught on quickly and is currently used as a general term for the use of Web 2.0 tools for e-learning (Ebner, 2007), such as the use of social networking sites to form communities of practice, learners who create their own (multimedia) content in the form of blogs, and the use of Wikipedia articles and YouTube videos as reference material.

Building upon the increasing popularity of Web 2.0 services for learning, so-called Personal Learning Environments (PLEs) are being developed (Gillet, 2010), environments that are composed of one or more (existing) subsystems – in many cases consisting of Web 2.0 services that are combined in a mash-up.

Apart from dedicated e-learning activities, learning often takes place in the context of our work or for private purposes. For this type of everyday learning, the organisation and retrieval of information items are essential activities. This will be briefly discussed in the next section.

3 Everyday e-learning at work and at home

Learning has become an integral part of many people's everyday life. A more knowledge-based society and rapid changes in technology require practically everyone to search for and read information in order to keep up-to-date. The character of learning has shifted from a solitary, paper-based activity to a web-based activity, making use of various resources, including discussion forums and social networking sites (Chatti and Jarke, 2007). This does not only yield for informal learning or learning at the workplace: students are increasingly expected to use these tools in their learning activities (New Media Consortium, 2012).

Also at home, search engines, travel planners, dictionaries and other online services have become essential for dealing with everyday tasks. News sites, portals, online games and streaming video are popular resources for information and entertainment. We communicate with our friends via email, social networking, forums, blogs and chat. As a result, one ends up with a large collection of scattered digital resources related to both our professional and personal lives (Razmerita et al., 2009).

The activities that people perform in order to acquire, organise, maintain, retrieve and use information items such as web pages for both everyday tasks and learning activities are commonly referred to as Personal Information Management[4] (Teevan et al., 2008). Search engines and social bookmarking sites play a major role in finding new information and services. For organising and re-finding these items, web provide history mechanisms such as URL auto-completion, the forward and back buttons, bookmarks and the history sidebar. However, this support is found to be suboptimal and skewed toward a small set of frequently visited resources (Obendorf et al., 2007).

For these reasons, the analysis and prediction of online browsing behaviour and re-visitation patterns has received much attention from the research community as well as from industry (Adar et al., 2008; Chierichetti et al., 2010; Kumar and Tomkins, 2010; Parameswaran et al., 2010; Tyler and Teevan, 2010). In the next section we discuss several findings that are also of relevance for the field of e-learning.

4 How users visit and revisit web content

There is a substantial body of research on how people organise and re-find information on the web. In the mid-nineties, Tauscher and Greenberg (1997) recognised the web as a 'recurrent system' that follows several regularities. The average probability of a page visit to be a revisit is estimated to be 58%. The majority of these revisits are covered by a small set of frequently used pages as well as recently used ones, mostly triggered by the browser's back button. These sets of Most Frequently Used pages (MFU) and Most Recently Used pages (MRU) both follow a power-law distribution.

These regularities have been confirmed in later studies (Obendorf et al., 2007; Adar et al., 2008; Tyler and Teevan, 2010). However, Obendorf et al. (2007) discovered through a client-side clickstream study, that individual browsing behaviour might be substantially different from the average numbers. For instance, the usage of multiple browser windows and tabs reduces the usage of the back button.

Further, the growing popularity of service-oriented sites – e.g. travel planners, online stores and social networking – changed the concept of *revisiting content* into *re-utilisation* of web applications. Although short-term revisits and visits to popular pages are well-supported by URL auto-completion (history lists and bookmarks are reported to be rarely used), long-term revisits typically involve elaborate researching and retracing.

Adar et al. (2008) provided more details on why users revisit pages. Apart from backtracking, short-term revisits involve the monitoring of news sites, as well as visits to shopping, search and reference websites. Long-term revisits involve specialised sites that are relevant every once in a while, and pertain to travel planning, job searching and weekend activities. Communication sites – web mail and forums – are represented in both categories. In a follow-up study, Tyler and Teevan (2010) analysed the use of search engines for re-finding. Results show that up to 39% of all queries involved re-finding; queries for re-finding are often used as a substitute for bookmarks. Still, less frequent revisits are not supported sufficiently enough, neither by search engines nor by web browser history mechanisms (Obendorf et al., 2007).

Due to the fact that e-learning has become a regular web activity, these findings should be taken into account when analysing learner behaviour in a LMS, as this will most likely bear similarities – in terms of revisits and parallel browsing – with the learner's browsing behaviour in other contexts. For instance, it has been observed that different kinds of sites invoke different kinds of behaviour (Obendorf et al., 2007). Search engines and reference sites have mainly one page – the portal – that users visit most often and a long tail of pages that users visit only once or twice. By contrast, LMS – as well as institutional and project websites – usually contain several locations that remain of interest to the learner (such as overview pages, reference pages, result summaries and discussion forums). Further, as interaction with LMS is often intertwined with 'regular' web use as well, one might miss important events when analysing the interaction with a LMS in isolation (Butoianu et al., 2010): for instance, at which point do learners often visit external resources and where do learners often leave a course and start pursuing spare-time activities.

5 Web usage mining for e-learning

Data mining in LMS is an emerging discipline, in which statistics, visualisation, classification, clustering and predictive methods are used for exploring and analysing the behaviour of learners (Romero et al., 2008; Graf et al., 2011). Typically, the basis for data mining is the LMS database, which may contain structured information about accesses to course material (assignments, choice, journal, lesson, quiz and survey) and student interactions (such as chat, forum, glossary, wiki and workshop). The data preparation for educational data mining can be summarised as follows:

- Selecting the relevant data.
- Summarising the data in a suitable format.

- Discretisation of numerical values.
- Transformation of the data to the required format for data mining or visualisation.

The more general field of web usage mining follows a similar approach (Mobasher, 2007). Most tools for web usage mining are designed and used for electronic commerce, with the goal to increase the number of visitors and visits, sales and profit. Even though the same techniques can be used for educational data mining, there is only limited knowledge on typical access patterns in e-learning contexts (Zaiane, 2002).

Further, many web usage algorithms – such as association rules – take as a building block a session or a purchase transaction. By contrast, in e-learning, learning sessions can span weeks or even months. A further difference is that in e-learning the goal of usage mining is less clearly defined. E-commerce sites can directly evaluate the effectiveness of the techniques by correlating them with an increase in sales or customer loyalty. The goals in e-learning are more generic and therefore difficult to directly qualify or quantify. Popular indicative measures for educational data mining include learning performance, drop-out rates and student motivation and satisfaction (Romero et al., 2008).

As has been noted before, learning does not take just place in a LMS or PLE, but in a heterogeneous (web) environment that includes both learning tools and services for collaboration and information management (Butoianu et al., 2010). Therefore, *learning analytics*, loosely defined as the combined use of web usage mining and educational data mining, is listed as a 'technology to watch' in the Horizon 2012 report (New Media Consortium, 2012).

It is beyond the scope of this paper to discuss web usage mining techniques in detail. More in-depth information or a general introduction to the field can be found in the work of Mobasher (2007) and Liu (2007). As a starting point for educational data mining or learning analytics, we refer to Romero et al. (2008) and Graf et al. (2011).

6 User data on the web and in e-learning

Using a web application often requires providing personal information to the system or to the service provider (Kobsa, 2001). The most explicit information is profile information (which includes name, demographics and other self-descriptions). In LMS, these profiles are typically called 'learner profiles'. Users typically can indicate which parts of the profile are public, shared or private.

Other information is explicitly given by the user, as a consequence of using this service: web email is stored at the email provider, GoogleDocs are stored at Google, Facebook messages are stored by Facebook, comments on a news article are stored on a news site and quiz results are stored in the LMS database. In many systems, users upload documents and write messages, comments or annotations themselves. Similar to profile information, users typically can indicate which documents, messages or comments are public, shared or private. Often this is arranged in default settings: Twitter messages are always public, Diigo annotations are public by default.

Finally, a lot of information is automatically gathered by the application. This includes search history, product browsing, learning activities and usage statistics. Even after having created a profile and having agreed with the terms and conditions, many users are not aware to what extent their actions are logged and what is done with this

information (Jernigan and Mistree, 2009). In most cases, this data is not made public, but is used for product recommendations, personalised search, history search, learning support or any other kinds of personalised services.

Typical e-learning-related activities that are logged by LMS include access to learning material, quizzes and assignments, grades, as well as comments, messages and notifications (Romero et al., 2008; Lonn et al., 2011). Apart from this learning-related data, it is often argued that this data should be *contextualised* by also taking into account the non-learning tools and services that are accessed during learning, such as e-mails, instant messaging and web pages (Butoianu et al., 2010).

6.1 Data reuse and synchronisation

More and more applications cooperate with other applications. One type of cooperation is realised through mash-ups, in which the user has integrated access to several applications. In e-learning, mash-ups are typically called PLEs (Drachsler et al., 2009). Mash-ups can be loosely connected or closely connected.

A different type of cooperation that becomes more and more popular is synchronising (syncing[5]). This can be a one-time process (importing your Facebook contacts into your email program, synchronising your Google calendar with your PDA) or it can be continuous (Facebook messages are also distributed in Twitter and vice versa). Syncing is a powerful mechanism for reducing user effort and providing better services.

To facilitate syncing, and for reducing the numbers of user names and passwords, OpenID[6] has become a (moderately) popular solution. The idea is that an OpenID provider maintains a central repository of your identities and that one can use just one username/password for all cooperating services. This does not mean that the OpenID provider has access to the applications of which it stores the user's profile information; for syncing one still needs to connect application to application.

The implication of the increasing use of mash-up and syncing techniques is that the data stemming from one application provides only a partial view on a user's activities. The data available in a LMS (such as visits to course pages and test scores) provides valuable knowledge, but in order to get the whole picture, one needs complementary information on the information accessed and messages communicated through other channels and applications. Correspondingly, the analysis of learner activities bears more and more similarities with 'regular' web usage mining.

7 Privacy concerns

The collection of user data is inherently connected with issues regarding online privacy. Online privacy is a topic that is increasingly covered by the media. Privacy denotes an individual's right to decide what information is made available to others (Westin, 1967). Control over personal information is essential in order to maintain relationships of varying degrees of intimacy, as desired by the individual.

The notion of online privacy adds the technological aspect to the legal and ethical aspects that surround privacy issues. This multi-dimensionality has made online privacy an issue of concern for various publics, including consumers, consumer advocacy groups, the media, marketers and governments.

The combination of online data from various channels has led to practices such as cyber-stalking, cyber-bullying or online identity theft. It is therefore crucial to equip users of all age groups with a sound understanding of what sharing data online can entail and to help them develop a stronger sense of responsibility for their own data. Similarly, web service providers and researchers should understand the risks of providing online user data to the general public.

In this section we explain why anonymisation is often not sufficient for addressing privacy issues and discuss techniques for ensuring privacy without rendering the user data unsuitable for analysis purposes.

7.1 Anonymisation is often not enough

Anonymising user data involves the removal of any references to the user's identity. This includes the users' names, login names, as well as IP addresses used by the users. The aftermath of the 2006 release of the AOL search data set[7] has shown that anonymisation is not sufficient for protecting the users' privacy. As reported in an article in the New York Times[8], triangulation of various queries of a user can be sufficient to track down the identity of a particular user.

More recently, Jones et al. (2008) discovered that in a sufficiently large query log, about 30% of users query for their own name. Even more, 50% of users issue queries that reference to their home location, postal code or ZIP code. They conclude that an attacker with access to a query log can realistically identify the names or locations of many users in it. Narayanan and Shmatikov (2008) demonstrated that only a small part of such data is needed to de-anonymise users of the Netflix movie data set.

For this reason, there is much interest in the problem of obscuring log data so that the privacy of individual users is protected, but the data remains useful for providing personalised services of for analysis. An extreme solution would be to encrypt all web locations visited and all queries issued. However, any detailed qualitative analysis or interpretation of the data will become hard or even infeasible in such an environment.

For the analysis of user activity within a LMS, with the aim to optimise its content and structure, information such as the learners' quiz results and the content visited before obtaining this result is essential (Romero et al., 2008). As another example, Bull and Kay (2005) discuss privacy concerns in the context of open learner modelling, and how these concerns may limit opportunities for collaboration and peer input. At the same time, this type of data is even more prone to reveal the identities of individual learners.

7.2 Techniques for ensuring privacy

There are several privacy-ensuring techniques that do not imply complete obscuration of meaningful but sensitive data. In this subsection we summarise the most relevant methods, which are explained in more detail by Kobsa (2007).

Berkovsky et al. (2005) propose the technique of *obfuscation*, which implies that a certain percentage of user actions become replaced by different values before being submitted to a central repository. In their approach, users are supposed to choose themselves which of their data should be obfuscated. This technique prevents that any conclusions drawn from the data cannot be adhered to one or more specific individuals with 100% certainty.

Perturbation is an approach that is similar to obfuscation. Perturbing data involves systematic alterations of the data – this may include that certain user actions are attributed to another user with more or less similar behaviour. This technique only has limited implications on the validity of analysis outcomes and still allows for qualitative interpretation.

A more privacy-preserving technique is the *aggregation* of data from multiple users, even though this comes at the cost that one cannot analyse individual user behaviour and differences between individual users.

Van Heerde et al. (2009) recognise the drawbacks of hiding identities or decoupling user actions from user profiles. They suggest a different solution which they call *data degradation*. After a predefined retention period, a privacy-preserving technique will be applied to make data less accurate and therefore at the same time less sensitive.

Which of the above techniques could or should be applied, depends on the sensitivity of the data and to what extent this data is expected to help in revealing individual users' identities; anonymisation involves a trade-off between privacy and fidelity. However, it has been shown that it is hard to distinguish between sensitive and non-sensitive data (Terrovitis et al., 2011). In particular, background knowledge may play an important role in retrieving a user's identity.

7.3 Summary

From the above discussion it has become clear that it is essential to deal with user data in a responsible manner. Researchers should carefully consider which data they need, which privacy-ensuring techniques are in order and whether it is prudent to release the data to other researchers or to the general public. Moreover, these decisions should be clearly communicated to prospective volunteers.

Together, knowledge and control are the fundamental prerequisites for users to give their informed consent to the collection of particular types of data. Data collection and dissemination on the web are unethical without obtaining users' informed consent beforehand. According to the Theory of Informed Consent, people can only consent to something, if they have received sufficient information, have understood it and have explicitly expressed agreement (Faden and Beauchamp, 1986).

8 Motivating participation

In order to motivate people to participate, one needs to convince prospective participants that your project is worth investing their time and effort – and potentially sacrificing some of their privacy. The most straightforward way to do so is by offering direct benefits, in most cases in terms of monetary compensation. Microsoft User Research and the Google User Experience Labs offer participants gift cards or other tokens of appreciation.[9,10] Apart from a monetary compensation – which may be regarded as mainly symbolic – they promise participants that they will get the opportunity to '[play] with some very cool technology and new games no one else has seen yet'.

Monetary compensation is also one of the main motivations for participants in the crowdsourcing internet marketplace, Amazon's Mechanical Turk.[11] In MTurk, tasks are distributed that computers are unable to do (yet). People can sign up to perform these

tasks, which are often quick to perform and for which the compensation is rather low. Distributing tasks to the 'crowd' is the Web 2.0 approach to *human computation*.[12] Results from a 2010 survey (Ipeirotis, 2010) indicate that apart from monetary compensation, motivations to participate include 'a fruitful way to spend free time', to 'kill time' and 'I find the tasks to be fun'.

Research on volunteerism in general suggests that if volunteers do not perceive benefits for themselves that they are less likely to start or continue volunteering (Panciera et al., 2011). This principle also yields for seemingly pure idealistic communities that aim to support a cause or to improve the current situation (Shen and Monge, 2011). For example, at first sight, it seems that the open source community is driven by people who share certain values and ideals, whose main goal is to provide direct benefits to their community. Indeed, active open source developers are often able to articulate how their contributions made a difference to a group effort. However, many large efforts – such as the Mozilla Firefox and Thunderbird project, various Linux distributions, the Apache Foundation – are supported by companies or foundations, with paid employees and a business model that relies on incomes that may be generated by advertisements, deals with commercial vendors or voluntary contributions (Midgley, 2006). But also smaller projects are not just driven by idealism, their own need for software or the enjoyment of creation: reputation in the community and career benefits are shown to be the major drivers for open source developers (Shen and Monge, 2011). For example, many extensions for the popular CMS Joomla are implemented by programmers who build on their portfolios.

Further, apart from active contributors and collaborators, the larger part of the open source community consists of *consumers*, the people who use the products and who provide feedback. According to Panciera et al. (2011), these consumers are less likely to feel involved in the community and therefore less inclined to contribute – by actively providing feedback or bug reports or even by allowing applications to collect (anonymous) usage data. Still, there are several cases of successful initiatives to collect user data, most notably the Firefox open data program, Test Pilot.[13] Currently, the program has over three million users, which allow for studies that achieve statistical significance without having to involve the whole population for each single study.

9 Web history repository

As discussed earlier in this paper, for the improvement and evaluation of concepts and tools for e-learning or personal information management in general, the availability of suitable data has become increasingly important. Companies like Google, Microsoft and Facebook constantly improve their products, based on the data that they collect from their users. Open-source software developers and researchers normally have no access to this data, which puts them into a disadvantage.

In this section, we discuss our experiences in building the WHR, which aims to address this disadvantage by creating a public repository of web usage data that researchers can use to gain new insights in online browsing behaviour. Using a Mozilla Firefox add-on, users can upload their anonymised usage data to the server.

9.1 Privacy concerns

Anonymity was a central issue during the conceptualisation of the WHR. As discussed in Section 7.1, the uproar around the AOL Query Log Data set in 2006[14] showed that it is not sufficient to hide the identity of the users; by combining evidence from search queries, it is still possible to trace these queries back to one specific person. As the WHR is meant to be available to the public, all URLs and hosts are encrypted using the MD5 algorithm. Apart from the factual risk that user identities will be involuntarily become exposed, the rigorous removal of any potential privacy concerns was also meant to minimise subjective worries about privacy loss.

Extensive information was provided on what information will be sent and how this is encrypted: 'The data you will submit to the Firefox WHR project is totally anonymised. No user identity is submitted nor the IP address. All the sensitive data – URLs, titles and site names – are encrypted and converted to Globally Unique Identifiers (GUID). The add-on does not track data – your anonymised data will only be sent once you click on the "Send now" button yourself'.

It turned out that obscured data is not always as obscured as it seems. We discovered that the social bookmarking service Delicious used the same encryption method for URLs as we did. As a result, we were able to use Delicious for obtaining tags that were related to the pages a user visited. As we only obtained this metadata and not specific URLs and queries, this procedure did not compromise the privacy of the contributors. However, in order to avoid potential privacy breaches, we currently distribute the WHR with double encrypted URLs.

9.2 Motivating users

The biggest challenge was (and still is) to convince regular users to contribute to the repository. It is unlikely that some user will go to the trouble of installing an add-on and uploading their highly personal data, just to make two researchers happy. For this reason, an explicit effort was made to communicate the importance of the WHR for researchers and open-source developers:

> "*Did you ever spend minutes or even hours trying to re-find a specific page? Do you want your Web browser to be smarter than just recommending the last visited pages or just showing you a list that you have to dig through yourself? So do we. And it is easy for you to help if you are a regular user of Firefox: just send us your anonymized Web history*".

Before the launch of the WHR project website, several rounds of user testing were held. An extensive review process by Mozilla volunteers provided a guarantee to the user that the add-on[15] is 'safe'. We promoted the WHR through several targeted mailing lists, Facebook, Blogspot and Twitter. After the first disappointing two weeks, during which we continued promoting the project, a critical threshold of 'fans' and 'followers' (including a number of influential people) was reached. Currently, a year after the release of the add-on, more than 400 anonymous volunteers contributed over four million entries from their browser history; every day 5–10 new contributions are added to the repository.

The data in the WHR includes the list of visited pages, including timestamp and browser session. For each visited page the (encrypted) URL and host, the total number of visits, the frequency and the last visit is listed in a separate table.[16] The WHR does not allow for extensive qualitative analysis as a similar data set that we gathered in 2006

(Weinreich et al., 2006), but the far larger quantity of data provides a good basis for analysis of *patterns* in online browsing behaviour as well as for evaluation of machine learning algorithms. Such analyses can lead to improved models of online browsing behaviour, better page (revisit) recommendations and mechanisms for organising and re-finding information and services found on the web.

9.3 Applications

We used the WHR for the evaluation and further improvement of SUPRA, a generic framework for real-time, contextual prediction of user navigation on the web (Kawase et al., 2010; Kawase et al., 2011a). The first layer of the framework calculates the a priori probability that a page will be revisited, based on its decency and frequency of use. The second layer complements the a priori probability with contextual methods that identify resources that are commonly visited within the current user context. We evaluated the benefits of the framework by integrating it in a dynamic browser toolbar, called 'PivotBar' that recommends visited pages that are relevant to the page currently viewed.

Further qualitative insights were gained by combining the WHR data with the tags provided by the Delicious bookmarking service – see Section 9.1. Delicious contained tags for only a subset of the pages in the web usage logs. Still, this subset sufficiently covered the long tails in the user's logs. In order to identify 'canonical' patterns of recurrent user interests, we followed the clustering and classification approach introduced by Adar et al. (2008).

The results indicate that the greater part of user interests involves tasks that turn up on a more or less regular basis and typically involve long-lasting activities as travel planning, project work and (goal-directed) shopping activities. In a nutshell, if an interest remains longer than one day, it is likely to return at a later stage (Kawase et al., 2011b). An example application of these findings is the selection and organisation of relevant reference material for learning or working contexts.

Both the WHR data set[17] and the contextual prediction framework SUPRA[18] are available to the community for further experimentation.

10 Discussion and conclusions

The fields of educational data mining and web usage mining bear many similarities. First of all, both fields involve the application of data mining and machine learning techniques on user data. But there are several more reasons to connect the two fields.

- E-learning has evolved into a 'regular' web browsing activity, which is often intertwined with 'normal' browsing.
- Learners often consult search engines and reference pages that are outside of the LMS.
- Mash-ups and synchronisation using Web 2.0 techniques become increasingly more common.

What sets educational mining apart from web usage mining, is the fact that – in addition to page access data – often other data is available and relevant as well, such as more detailed learner profile data and specific quiz results. A further difference is that learning activities invoke different behaviour than 'regular' web browsing.

In this paper, we discussed these similarities and differences and how the two fields contribute to the emerging field of learning analytics. In particular, we focused on the need for data sets and considerations to be taken if one aims to recruit participants for creating such a data set. We used the WHR as an example of a successful approach in order to provide concrete pointers and guidelines.

References

Adar, E., Teevan, J. and Dumais, S.T. (2008) 'Large scale analysis of web revisitation patterns', *Proceedings of CHI*, 5–10 April, Florence, Italy, pp.1197–1206.

Berkovsky, S., Eytani, Y., Kuflik, T. and Ricci, F. (2005) 'Privacy-enhanced colalborative filtering', *UM05 Workshop on Privacy-Enhanced Personalization*, 25 July, Edinburgh, UK.

Bull, S. and Kay, J. (2005) 'A framework for designing and analysing open learner modelling', *Proceedings of Workshop on Learner Modelling for Reflection*, Amsterdam, The Netherlands, pp.81–90.

Butoianu, V., Vidal, P., Verbert, K., Duval, E. and Broisin, J. (2010) 'User context and personalized learning: a federation of contextualized attention metadata', *Journal of Universal Computer Science*, Vol. 16, pp.2252–2271.

Chatti, M.A. and Jarke, M. (2007) 'The future of e-learning: a shift to knowledge networking and social software', *International Journal of Knowledge and Learning*, Vol. 3, Nos. 4/5, pp.404–420.

Chierichetti, F., Kumar, R. and Tomkins, A. (2010) 'Stochastic models for tabbed browsing', *Proceedings of 19th International Conference on World Wide Web*, 26–30 April, Raleigh, NC, USA, pp.241–250.

Downes, S. (2005, October) 'E-learning 2.0', *eLearn Magazine*.

Drachsler, H., Rutledge, L., Van Rosmalen, P., Hummel, H.G.K., Pecceu, D., Arts, T., Hutten, E. and Koper, R. (2009) 'Remashed – an usability study of a recommender system for mash-ups for learning (special issue)', *Proceedings of ICL*, 23–25 September, Villach, Austria.

Ebner, M. (2007) 'E-Learning 2.0 = e-Learning 1.0 + Web 2.0?', *Availability, Reliability and Security,* 10–13 April, Vienna, pp.1235–1239.

Faden, R.R. and Beauchamp, T.L. (1986) *A History and Theory of Informed Consent,* Oxford University Press.

Gillet, D., Law, E.L-C. and Chatterjee, A. (2010) 'Personal learning environments in a global higher engineering education web 2.0 realm', *IEEE EDUCON Education Engineering 2010 Conference*,14–16 April, Madrid, pp.897–906.

Graf, S., Ives, C., Rahman, N. and Ferri, A. (2011) 'AAT – a tool for accessing and analysing students' behaviour data in learning systems', *Proceedings of 1st International Conference on Learning Analytics and Knowledge, LAK 2011*, 27 February–1 March, Banff, AB, Canada, pp.174–179.

Ipeirotis, P.G. (2010) *Demographics of Mechanical Turk*, Technical Report CEDER-10-01, New York University.

Jernigan, C. and Mistree, B.F.T. (2009) 'Gaydar: Facebook friendships expose sexual orientation', *First Monday*, Vol. 14, No. 10.

Jones, R., Kumar, R., Pang, B. and Tomkins, A. (2008) 'Vanity fair: privacy in querylog bundles', *Proceedings of the 17th ACM Conference on Information and Knowledge Management* CIKM 2008, 26–30 October, Napa Valley, California, USA, pp.853–862.

Kawase, R., Papadakis, G. and Herder, E. (2010) 'How predictable are you? A comparison of prediction algorithms for web page revisitation', *18th International Workshop on Personalisation and Recommendation on the Web and Beyond (ABIS)*, Kassel, Germany.

Kawase, R., Papadakis, G. and Herder, E. (2011a) 'Beyond the usual suspects: context-aware revisitation support', *Proceedings of the 22nd ACM conference on Hypertext and Hypermedia HT'11*, 6–9 June, Eindhoven, The Netherlands, pp.27–36.

Kawase, R., Papadakis, G. and Herder, E. (2011b) 'Supporting revisitation with contextual suggestions', *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL*, 13–17 June, Ottawa, ON, Canada, pp.227–230.

Kobsa, A. (2001) 'Generic user modeling systems', *User Modeling and User-Adapted Interaction*, Vol. 11, pp.49–63.

Kobsa, A. (2007) *Privacy-Enhanced Web Personalization*, Springer Verlag, pp.628–670.

Kumar, R. and Tomkins, A. (2010) 'A characterization of online browsing behavior', *Proceedings of the 19th international conference on World Wide Web WWW'10*, 26–30 April, North Carolina, USA, pp.561–570.

Liu, B. (2007) *Web Data Mining – Exploring Hyperlinks, Contents, and Usage Data*, Springer.

Lonn, S., Teasley, S.D. and Krumm, A.E. (2011) 'Who needs to do what where?: Using learning management systems on residential vs. commuter campuses', *Computers and Education*, Vol. 56, pp.642–649.

Midgley, S. (2006) 'The case for open markets in education', *First Monday*, Vol. 11, No. 7.

Mobasher, B. (2007) *Data Mining for Web Personalization*, Springer Verlag, pp.90–135.

Narayanan, A. and Shmatikov, V. (2008) 'Robust de-anonymization of large sparse datasets', *IEEE Symposium on Security and Privacy*, 18–22 May, Oakland, CA, pp.111–125.

New Media Consortium (2012) *Horizon Report – 2012 Higher Education Edition*, New Media Consortium.

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', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI'07*, 28 April–3 May, San Jose, California, USA, pp.597–606.

Panciera, K., Masli, M. and Terveen, L. (2011) 'How should i go from – to – without getting killed? Motivation and benefits in open collaboration', *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym'11*, 3–5 October, Mountain View, California, pp.183–192.

Parameswaran, A.G., Koutrika, G., Bercovitz, B. and Garcia-Molina, H. (2010) 'Recsplorer: recommendation algorithms based on precedence mining', *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 6–10 June, Indianapolis, Indiana, USA, pp.87–98.

Razmerita, L., Kirchner, K. and Sudzina, F. (2009) 'Personal knowledge management – the role of web 2.0 tools for managing knowledge at individual and organisational levels', *Online Information Review*, Vol. 33, pp.1021–1039.

Romero, C., Ventura, S. and Garcia, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', *Computers and Education*, Vol. 51, pp.36–384.

Shen, C. and Monge, P. (2011) 'Who connects to whom? A social network analysis of an online open source software community', *First Monday*, Vol. 16, No. 6.

Tauscher, L. and Greenberg, S. (1997) 'How people revisit web pages: empirical findings and implications for the design of history systems', *International Journal of Human-Computer Studies*, Vol. 47, No. 1, pp.97–137.

Teevan, J., Jones, W. and Capra, R. (2008) 'Personal information management (PIM) 2008', *Sigir Forum*, Vol. 42, No. 2, pp.96–103.

Terrovitis, M., Liagouris, J., Mamoulis, N. and Skiadopoulos, S. (2011) *Privacy Preservation by Disassociation*, Technical Report TR-IMIS-2011-1, Institute for the Management of Information Systems, Athena RC, Greece.

Tyler, S.K. and Teevan, J. (2010) 'Large scale query log analysis of re-finding', *Proceedings of the 3rd International Conference on Web Search and Web Data Mining WSDM 2010*, 4–6 February, New York, NY, USA, pp.191–200.

Van Heerde, H., Fokkinga, M.M. and Anciaux, N. (2009) 'A framework to balance privacy and data usability using data degradation', *IEEE International Conference on Computational Science and Engineering*, 29–31 August, Vancouver, BC, Canada, pp.146–153.

Weinreich, H., Obendorf, H., Herder, E. and Mayer, M. (2006) 'Off the beaten tracks: exploring three aspects of web navigation', *Proceedings of the 15th International Conference on World Wide Web WWW'06*, 22–26 May, Edinburgh, Scotland, UK, pp.133–142.

Westin, A.F. (1967) *Privacy and Freedom*, Atheneum.

Zaiane, O.R. (2002) 'Building a recommender agent for e-learning systems', Proceedings of International Conference on Computers in Education, ICCE, 3–6 December, Auckland, New Zealand, Vol. 1, pp.55–59.

Notes

1 http://moodle.org/

2 http://www.blackboard.com/

3 http://masieweb.com/Surveys/learning-systems-survey-results.htm

4 http://en.wikipedia.org/wiki/Personal information management

5 http://en.wikipedia.org/wiki/File synchronization

6 http://openid.net/

7 http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

8 http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all

9 http://www.microsoft.com/userresearch/enrollWho.aspx

10 http://www.google.com/forms/user_faq.html

11 https://www.mturk.com/mturk/welcome

12 http://en.wikipedia.org/wiki/Human-based_computation

13 http://design-challenge.mozillalabs.com/open-data/OpenDataCompetition/

14 http://techcrunch.com/2006/08/07/aol-this-was-a-screw-up/

15 https://addons.mozilla.org/en-US/_refox/addon/226419/

16 https://wiki.mozilla.org/Places:Design_Overview

17 http://webhistoryproject.blogspot.com/

18 http://sourceforge.net/projects/supraproject/