Visualising web server logs for a Web 1.0 audience using Web 2.0 technologies: eliciting attributes for recommendation and profiling systems

Ed de Quincey[1], Patty Kostkova[1] and David Farrell[1]

[1] City eHealth Research Centre (CeRC), City University, London, UK
{Ed.de.Quincey}@city.ac.uk

Abstract. Web server logs have been used via techniques such as user profiling and recommendation systems to improve user experience on websites. The data contained within server logs however has generally been inaccessible to non-technical stakeholders on website development projects due to the terminology and presentation used. We describe a process that uses visualisation to enable these stakeholders to identify questions about site usage including user profiling and behaviour. The development of this tool utilising Web 2.0 technologies is described as well as feedback from the first stage of user evaluation on a real-world multi-national web development project called e-Bug. The potential for this process to elicit user attributes and behaviour that can be incorporated into automated user profiling systems is also discussed.

Keywords: Visualisation, Web Server Logs, User Profiling, Web 2.0 Technologies

1 Introduction

Research into online user behaviour has been aided by the relative ease of collecting feedback data using implicit methods such as web server logs [1, 2, 3], compared to explicit methods such as usability testing [4, 5], tagging [6] and ratings [7]. The data stored in server logs has been used to create a number of recommendation [8, 9, 10] [11, 12] and profiling systems [13, 14].

This has had a dramatic impact on the user experience e.g. Amazon [15] but apart from deliberate or accidental releases of server log data (e.g. NetFlix Prize[1], AOL), the information contained within the logs has been generally hidden from the users of a website and more importantly from non-technical stakeholders of a web development project. This means that few people outside of the server log analysis or web development communities fully understand the information that is stored in web logs and the user behaviour that it can explain.

[1] http://www.netflixprize.com/

There have been several commercial attempts (Google Analytics[2], Sawmill[3], WebTrans[4]) that have tried to make server logs, and therefore user behaviour, more accessible to site owners. However, these applications analyse generic features of sites that do not answer specific questions that certain stakeholders will have and do not help them identify trends in user behaviour due to the sheer volume and technical nature of the information presented [16].

A potential solution to this problem is to use techniques from the field of Software Visualisation (SV) to make the data contained within server logs more accessible to non-technical stakeholders in website development projects. Using these methods utilises the innate pattern matching ability [17] of the human cognitive system to identify trends in user behaviour which might be missed by the current automated profiling and recommendation systems. Once identified, non-technical stakeholders, such as content providers, can adapt content and the site design to fit user behaviour [16]. This human expertise could then potentially be integrated into current automated recommendation and profiling systems.

This paper describes the process of developing and using visualisation techniques to disseminate site usage information to non-technical stakeholders, in order to identify potential attributes for user profiling and recommendation systems. An ongoing multinational project in e-Health, called e-Bug (www.e-bug.eu), has been used as a test-bed and feedback from project stakeholders is detailed. The future possibilities of this technique are discussed as well as general implementation issues from using Web 2.0 technologies.

2 Background Information

2.1 Visualisation and Metaphors

Visualisation is concerned with making large amounts of information more comprehensible for the user by using a visual representation. Software Visualisation has been successfully used by software engineers to "make software more visible" [18] by representing the significant features of code using a visual metaphor. A well known example of a visualisation is the London Underground Tube Map[5] which is a representation of a complex, real world artifact that can be understood immediately and navigated simply. A detailed taxonomy of SV has been produced by Brice et al. [19] and also the related fields of Information Visualisation, Visual Analytics [20][21] and Metaphors used in interface design [22] contain a number of related and relevant techniques.

[2] http://www.google.com/analytics/
[3] http://www.sawmill.net/
[4] http://www.webtrans.co.uk/
[5] http://www.tfl.gov.uk/gettingaround/1106.aspx

2.2 The e-Bug Project

e-Bug is a European Commission funded project that aims to reduce inappropriate antibiotic use and improve hygiene through improving the education of young people in seventeen participating countries. e-Bug combines traditional methods of classroom delivery with online, browser-based (Flash) games to teach a pupils in junior and senior schools about microbes, hand and respiratory hygiene, and antibiotics. Example lessons and media are available on the e-Bug website[6] alongside games that can be used alongside the pack or standalone [23].

Currently the server logs from the e-Bug project are analysed using a proprietary application called Sawmill. This produces standard reports that cover information such as visits, hits, content viewed, visitor demographics and systems and referrers. These reports are produced monthly and uploaded onto the e-Bug website[7]. It was found however that although the project partners expressed a high degree of interest in the website statistics during meetings, the format that the reports were currently in were not easily accessible to non-technical users. This was mainly due to the terminology used and the statistics presented not answering specific questions that the project partners had regarding the users of the site [D. Farrell 2009, pers. comm.]. It was decided therefore that the server logs from the e-Bug project website would make a suitable test-bed to use visualisation techniques to analyse and present the statistics in a way that reduced the confusion and elicited potential attributes for user profiling.

3 Method for server log visualisation

A User Centred Methodology (UCD) [24] was used to develop a prototype application that would visualise the statistics that were currently calculated by the Sawmill application e.g. visits during particular months/years, geolocations of visits.

Sketching has been used previously to create code visualisation software [25] and so the same approach was used initially to explore potential metaphors and representations that could be used. An example sketch is shown below in Figure 1.

[6] http://www.e-bug.eu
[7] http://www.e-bug.eu/ebug_secret.nsf/England-Project-General/eng_eng_p_wp_gn_stats

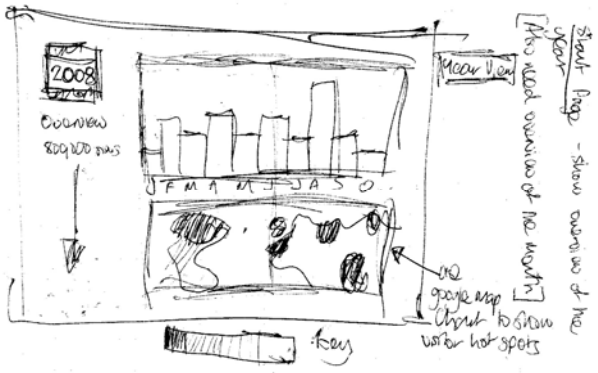

Fig. 1. Example sketch illustrating the weather map metaphor and bar charts

At this stage two potential metaphors were identified: a weather map metaphor and a timeline metaphor. After discussion with members of the project team it was decided to begin by developing the weather map metaphor as this would support one of the main features that was missing from the current reports: accurate geographical distribution of the users of the site.

3.1 Web 2.0 Technologies for Visualisation

Having identified possible interface designs for the application, an online prototype system was developed and suitable technologies explored for creating the map metaphor. The following figure shows the first version of the prototype.

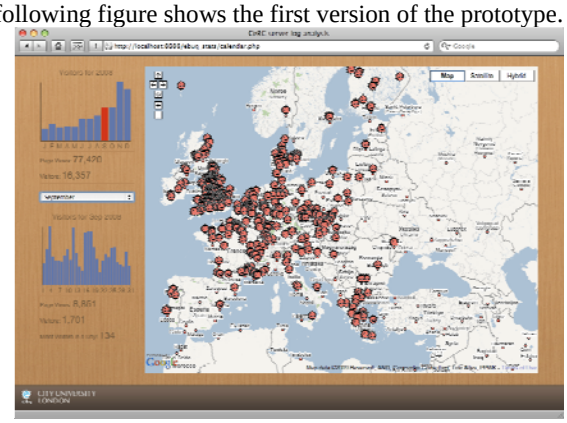

Fig. 2. Visualisation of visitors in September 2008 with each red icon representing a visitor

The interface incorporates two main visualisations. An area on the left hand side of the screen that shows the number of visitors and page views in a particular year and

month (and their daily distribution) using simple bar charts. The area on the right contains a map with individual visitors denoted by their location with a marker (in this case the e-Bug logo). Users can select particular months from the drop down menu on the left and navigate the map using the navigation icons and the mouse pointer.

The map was created using the Google Maps API, which uses JavaScript to make asynchronous calls (AJAX) to display the map and the markers. The data for the markers is stored in an XML file that is generated by a PHP page parsing a CSV file that is created using Sawmill[8]. The CSV file contains paired values of a users' hostname and the number of page views that came from that IP address. PHP is then used along with the GEOIP Lite Open Source reverse geolocation database[9] to calculate a longitude and latitude for each hostname. These are then saved in an XML file in the following format:

```
<marker lat="40.6333" lng="-7.8333"/>
```

The bar charts were created using the Google Charts API, which creates dynamic images based on parameters passed in the querystring, for example:

```
<img
src="http://chart.apis.google.com/chart?cht=p3&chd=t:60
,40&chs=250x100&chl=Hello|World" alt="Bar chart" />
```

The parameters were determined using PHP pages and CSV files that contain monthly and daily totals of visits and page views.

4 Evaluation

This prototype was then uploaded to the e-Bug website and feedback was elicited from members of the e-Bug project team from seventeen European countries, as well as researchers involved in similar projects at UK Universities as part of the UCD process. The evaluation was in the form of an email with a set of open-ended questions that respondents were asked to answer regarding the interface. The main focus of this exercise was to ascertain whether the information that was being represented was clear enough, whether appropriate metaphors were being used and also whether there were any other statistics that users would be interested in. As this is an ongoing project, feedback has so far been received from nine respondents.

The majority of respondents reacted positively to the interface and the visualisation and a number of them were able to give detailed feedback, indicating that they were able to understand what the page was showing and what it did not. The main recurring points from this feedback are detailed below:

- Add representation that shows "magnitude of visitors" as it is difficult to gauge repeat visitors, number of pages viewed and markers that overlap.

[8] The data from Sawmill was used rather than the raw server logs due to the fact that Sawmill filters out certain web crawlers as well as using custom filters that have been created to remove certain IP addresses.

[9] http://www.maxmind.com/app/geolitecity

- Add specific place markers to the map that do not appear (unless at a higher zoom level).
- Add specific evaluation areas/overlays onto the map[10].
- Show the density of visitors in each area i.e. show visitors per 100,000 population to get more meaningful comparisons.
- Add in a view of popular pages downloads and where they originate from.
- Highlight returning visitors.
- Add in a view that shows the times of day that various pages are being accessed e.g. if the games are being viewed outside of school hours this could indicate that students are playing them at home.
- Ability to compare months and countries.

One of the most interesting points noted by the stakeholders however was the fact that the data being represented itself is a potential area of confusion. For example, a number of users gave the general impression that they did not know the difference between a visitor and a hit. It became clear that the target users of this application do not posses the same knowledge that experts in the field take for granted and further investigation into this area is being conducted.

Following on from this, a second version is currently being developed to take into account the feedback and also to tackle some of the issues that have been raised with regards to the interface and the information that users would like displayed. A screenshot(s) from the second iteration of the software is shown below:

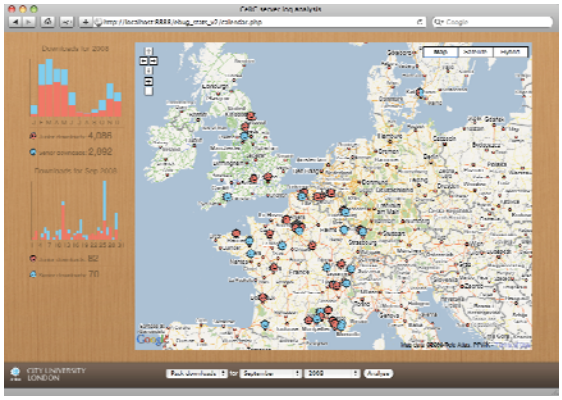

Fig. 3. Version 2 of the software visualises different types of file downloads, represented with two different colours

As well as markers and statistics for visitors, information regarding pack downloads (educational resources for teachers in Word and PowerPoint files) has been included and split into "Junior" and "Senior" versions.

This version of the application also uses an updated visitors' visualisation that takes into account the number of page views from a particular users. The well-known temperature scale visualisation used on weather maps has been utilised to be able to differentiate between the levels of activity in various regions.

[10] this can potentially be achieved using the Google Polylines' API

5 Discussion

5.1 Potential for use in User Profiling

Initial feedback has already indicated that visual representations of the data have allowed the non-technical stakeholders in the project to start to identify user types and user behaviour. One particular interest is whether pupils are accessing the games pages at home or at school and whether the tool can identify whether it is a student viewing the website or a teacher. By geographically representing visitors in relation to the location of target schools, along with the time they are accessing the site can potentially achieve this simple user profiling task.

This example and others detailed in Section 4 indicate that providing non-technical stake holders with a visual representation of the server logs has allowed them to communicate requirements for further analysis which can either be integrated into the filters used in the Sawmill application or into the visualisation tool. Without the use of visualisation techniques, it is doubtful that these questions regarding the users of the site and their behaviour would have been raised.

Further investigation of user profiles and understanding of national profiling differences is a subject of our ongoing research.

5.2 Strengths of Web 2.0 Technologies for Visualisation

There are a number of advantages with using Web 2.0 technologies such as the various Google API's and AJAX such as being able to create richer and more interactive online interfaces but the main advantage relates to being able to utilise users' pre-existing skills and experience. The majority of users have prior experience with interfaces such as Google Maps and in the same way that the Desktop has become the standard metaphor used for operating systems, maps and markers and the various methods of interaction that Google has developed have become a standard in this area. Being able to "piggy-back" on to that frees the user from the interface and allows them to focus on the visualisation, even though this application is a bespoke solution.

An associated advantage is that Google is a global organisation and so is its software. The potential users of this software are from a diverse set of countries with a number of different languages and levels of expertise. With Google being even more popular in Europe than the US [26], and its projected market share expected to take over the number one position from MapQuest by the end of the year [27], means that the chances of a user having had previous exposure to the Google Maps interface, and therefore the interface of this application, is quite high. This also has follow on advantages for issues such as localisation and internationalisation.

The other advantage is the increased speed in development. Being able to harness pre-existing API's allows for rapid prototyping and the ability to demonstrate a working concept to users to elicit feedback almost immediately and also allows for faster changes and incremental versions.

Finally, the fact that Web 2.0 technologies are also designed to be accessible via a number of different browsers and platforms also allows for speedier access and dissemination of the information which is vital for cross-nation projects such as e-Bug.

5.3 Limitations of Web 2.0 Technologies for Visualisation

One of the main problems with the Google Maps marker metaphor is the problem of occlusion, something that is common when using 3-D visualisations. If a user visits the site numerous times or downloads numerous pack or pages it is difficult to represent that with numerous markers on the map as they will overlap with one another. This can partially solved with the colour coding of markers but the accuracy of the geolocation database and the fact that numerous visitors can originate form the same area means that the markers often overlap. To improve this a method for clustering the markers so that close "neighbours" are represented by one marker and for this information to be presented textually once a user clicks on a clustered marker are being investigated.

A related problem is the amount of data that can be represented using these tools and the limitations of the browser. During testing of the application it was found that once around five thousand markers were placed on the screen using the standard Google method, the browser would slow down and become unusable. For this prototype this problem was solved by filtering out duplicate markers and also non-European hits (as this was not required at this stage in the site's development). However once the site it is launched and publicised further this year, there will be an increase in visitors and therefore an increase in markers. Clustering methods are therefore currently being investigated.

One final problem that was highlighted from user feedback was that relying on users having had prior experience on Google Maps means that for those who have not, or those who do not realise that this is a Google Maps interface, have initial problems with the interface. Adding extra methods of navigation or instructional video/instructions are currently being piloted.

6 Conclusion

The process of identifying appropriate visualisations to allow non-technical users to start to identify site usage from server logs is important for successful web site development and evaluation. The process presented in this paper has provided a number of insights into the potential of using Web 2.0 tools and metaphors for visualisation and dissemination of information. Although at an early stage, the tool is already providing insights into a number of usage patterns on the site which are enabling non-technical stakeholders of the e-Bug project to start to identify distinct user profiles and most importantly to start to be able to utilise the data stored in server logs more readily.

Future work will include an investigation into pre-existing taxonomies that exist of software visualisation [19] to see which might be relevant for representing web server

log data and also which can be supported by Web 2.0 technologies. Also, current visualisation techniques from the biological sciences will be studied to see if any of these are appropriate e.g. spread of user activity being represented in a similar way to disease spread.

Following on from this, the tool will be used in an investigation into user behaviour on the e-Bug website in order to see whether researchers can identify usage trends visually and what are the attributes of these trends e.g. time of day a user visits plus geographical location might indicate whether they are a pupil or a teacher. This will then feed directly into the development and tailoring of content for the site and the potential for incorporating this into an automated profiling system will be investigated.

References

1. Dupret, G. E. and Piwowarski, B.: A user browsing model to predict search engine click data from past observations. In Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Singapore, Singapore, July 20 - 24, 2008). SIGIR '08. ACM, New York, NY, 331-338 (2008)
2. Joachims, T.: Optimizing search engines using clickthrough data. In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM, New York, NY, 133-142 (2002)
3. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. ACM Trans. Inf. Syst. 25, 2 (Apr. 2007), 7 (2007)
4. Nielsen, J.: Usability inspection methods. In Conference Companion on Human Factors in Computing Systems(Boston, Massachusetts, United States, April 24 - 28, 1994). C. Plaisant, Ed. CHI '94. ACM, New York, NY, 413-414 (1994)
5. Spool, J. and Schroeder, W.: Testing web sites: five users is nowhere near enough. In CHI '01 Extended Abstracts on Human Factors in Computing Systems (Seattle, Washington, March 31 - April 05, 2001). CHI '01. ACM, New York, NY, 285-286 (2001)
6. Kipp, M. E. and Campbell, D. G.: Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. Available from http://dlist.sir.arizona.edu/1704/01/KippCampbellASIST.pdf (2006)
7. Anand, S. S., Kearney, P., and Shapcott, M.: Generating semantically enriched user profiles for Web personalization.ACM Trans. Interet Technol. 7, 4 (Oct. 2007), 22 (2007)
8. Schafer, J. B., Konstan, J., and Riedi, J.: Recommender systems in e-commerce. In Proceedings of the 1st ACM Conference on Electronic Commerce (Denver, Colorado, United States, November 03 - 05, 1999). EC '99. ACM, New York, NY, 158-166 (1999)
9. Schafer, J. B., Konstan, J. A. and Riedl, J.: Ecommerce recommendation application, Data Mining and Knowledge Discovery, 5(1/2):115–153 (2001)
10. Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J.T.: Item-based collaborative filtering recommendation algorithms, Proc. of the Tenth Int. WWW Conf., pp. 285–295 (2001)
11. Ali, K. and van Stam, W.: TiVo: making show recommendations using a distributed collaborative filtering architecture. In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Seattle, WA, USA, August 22 - 25, 2004). KDD '04. ACM Press, New York, NY, 394-401 (2004)

12. Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 1 (Jan. 2004), 5-53 (2004)
13. Kostkova, P., Diallo, G. and Jawaheer, G.: User Profiling for Semantic Browsing in Medical Digital Libraries, Proceedings of ESWC 2008, The Semantic Web: Research and Applications, 5th European Semantic Web Conference, Tenerife, Canary Islands, Spain, June 1-5 (2008)
14. Danilowicz, C. and Indyka-Piasecka, A.: Dynamic User Profiles Based on Boolean Formulas, Lecture Notes in Computer Science, Volume 3029/2004, pages 779-787 (2004)
15. Linden, G., Smith, B. and York, J.: Amazon.com Recommendations Item-to-Item Collaborative Filtering, IEEE Internet Computing, January-February (2003)
16. Drott, M. C.: Using Web server logs to improve site design. In Proceedings of the 16th Annual international Conference on Computer Documentation (Quebec, Quebec, Canada, September 24 - 26, 1998). SIGDOC '98. ACM, New York, NY, 43-50 (1998)
17. Sinha, P.: Recognizing complex patterns. Nat Neurosci 5:1093–1097 (2002)
18. Petre, M. and de Quincey, E.: A gentle overview of software visualisation. Computer Society of India Communications. August, 6-11. ISSN 0970-647X (2006)
19. Price, B. A. Baecker, R. M. and Small, I. S. A Principled Taxonomy of Software Visualization, Journal of Visual Languages and Computing, 4, 1993, 211—266 (1993)
20. Thomas, J.J. and Cook, K.A.: Illuminating the Path: The R&D Agenda for Visual Analytics. National Visualization and Analytics Center. p.3-33 (2005)
21. Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press (1983)
22. Erickson, T.: Working With Interface Metaphors, 65-73, edited by B. Laurel, The Art of Human-Computer Interface Design, AddisonWesley Publishing Company Inc, Reading, Massachusetts, USA. (1990)
23. Kostkova, P. and Farrell, D.: e-Bug online Games for Children: Educational Games Teaching Microbes, Hand and Respiratory Hygiene and Prudent Antibiotics Use in Schools across Europe, In the Proceedings of the ESCAIDE 2008 conference, Berlin, November 2008 (2008)
24. U.S. Department of Health & Human Services: Step-by-Step Usability Guide. Available: http://www.usability.gov/. Last accessed 4th February 2009. (2009)
25. Craft, B. & Cairns, P.: Using Sketching to Aid the Collaborative Design of Information Visualisation Software-A Case Study, In Human Work Interaction Design: Designing for Human Work, Springer Boston, Volume (221), p. 103—122 (2006)
26. Lipsman, A.: Google Holds Top Spot in European Site Rankings, According to comScore World Metrix. Available: http://www.comscore.com/press/release.asp?press=988. Last accessed 1st March 2009. (2006)
27. ABI Research: The Google Maps vs MapQuest Online Mapping Portal War Is Driving Map 2.0 Innovations. Available: http://tinyurl.com/dlwj7d. Last accessed 5th March 2009. (2009)