From User-Centric Web Traffic Data to Usage Data

Thomas BEAUVISAGE, Houssem ASSADI

France Telecom R&D
38-40, rue du Général Leclerc
92794 Issy les Moulineaux Cedex 9 - France
+33 (0) 1 45 29 58 11

{thomas.beauvisage, houssem.assadi}@francetelecom.com

ABSTRACT

In this paper, we describe a user-centric Internet usage data processing platform. Raw usage data is collected using a software probe installed on a panel of Internet users' workstations. It is then processed by our platform. The transformation of raw usage data into qualified and usable information by Internet usage sociology researchers means setting up a series of relatively complex processes using quite a wide variety of resources. We use a combination of ad hoc rule-based systems and external resources to qualify the visited Web pages. We also implemented topological and temporal indicators in order to describe the dynamics of Web sessions.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Sociology.
H.3.4 [Information Storage and Retrieval]: Systems and software.

General Terms: Design, Measurement, Human Factors

Keywords: User-centric traffic data, Web Usage Mining, traffic analysis, usage data, Internet uses.

1. INTRODUCTION

The present article describes a user-centric Internet usage data platform, which goal is to enable fine-grained user-centred analysis of Web uses. It has been developed as part of the cooperative project SensNet – a partnership between France Telecom R&D, NetRatings, CNRS/LIMSI and the University of Paris III; with a financial support from the French Ministry of Industry. In this project, we used this platform with data from Internet panels in France observed over 3 years (2000-2002).

2. COLLECTING USAGE DATA

Many systems have already been developed in the domain of Web Usage Mining for analysing Web server logs; however, this server-centric point of view is not the most relevant for a fined-grained description of Internet usage. Although more relevant (see for example the results from [5] and [7]), user-centric traffic data are more rare and more complex, considering the variety of sites, contents and usage contexts a user can encounter.

As part of the previously mentioned SensNet project, we used user-centric data collected by NetMeter, an unobtrusive probe developed by NetValue (this company and the technology mentioned are currently the property of NetRatings – see www.netratings.com).

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

Internet activity is monitored in real time using a software probe installed on each panellist's computer. The information is analysed by identifying the different users in the household. The probe records data about the most common Internet protocols (HTTP, NNTP, SMTP, POP3…). In the case of HTTP protocol, the probe records a time-stamped list of the URLs requested by the user, with the Referer and the page volume information.

3. DATA PREPARATION

The Web traffic data as described previously – roughly a list of URLs – is still in a very "raw" format and needs to undergo a series of processes to be actually usable. Data preparation stage raises a series of non trivial problems, even before any usage analysis is envisaged. We will not explain here in details the process of data preparation, but will mention two main steps:

1. Identifying sessions. This involves the need to recognise coherent sequences of user activity within continuous time-stamped traffic data. This recognition has already been the subject of a lot of works analysing Web server logs (see for example [6]). The important point here is to adopt the user's point of view, i.e. to include all traced protocols for identifying Internet sessions (Web, e-mail, chat…). Then, the difficulty is to set the relevant period of inactivity between two recorded events to determine whether a session is finished. Following a statistical study, this period has been set to 30 minutes.

2. Identifying sites. Using the URL's *host* field raises two main problems: reduction (ex : www.cpan.org considered as different from search.cpan.org), aggregation (personal Web sites hosted on the same DNS, like perso.wanadoo.fr). To face these problems, we propose the concept of the *editorial site*, considering a site to be a publication area with a single editorial entity. We developed methods to identify these *editorial sites*, based on 1) personal Web sites recognition and adapted segmentation heuristics, and 2) generic TLD (.com, .net…) and country code TLD (.fr, .po…) sub-domains pattern recognition. These methods allow us to have a reliable base for Web sites identification.

4. URL LEVEL ENRICHMENT

4.1 *CatService* module

The *CatService* module, in the SensNet platform, provides a qualification of the URLs visited in terms of types of site and service. There are five levels of qualification: site: s*ite type* (for example: "generalist portal", "WebMail" site, "digital library"), *site* or *portal* (ex: Yahoo), *service* and *sub-service* (ex: "search service" / "image search"), and *service supplier* (ex: Yahoo search service provided by Google). Each service and sub-service is related to a

site type. *CatService* users must define these categories, and can easily edit and modify them (defining site types, adding sites or services to a site type, etc.).

To operate, *CatService* requires a set of *pattern matching* rules, built with the help of a formalism based on regular expressions. These rules enable us to associate a class of URLs with a given portal-supplier couple, a service and a sub-service. These rules are constructed manually after examining the different addresses in a portal and verifying the page content to which they point.

Categorising services is of great value to our work, in particular for generalist portals. *CatService*'s detailed description not only distinguishes, in each portal's audience, between the different services used (search engine, WebMail, etc.), but also makes these elements comparable between different portals. It also introduces an important services concept and enables a distinction to be made between pages whose textual content takes precedence and those where the function (the service proposed) is more important from a descriptive point of view (ex: Information pages, or Communication services such as Webmail). In addition, *CatService* allows the study of specific services at a more detailed level: although it does not aim to cover all browsing, it enables us to select categories of sites which we want to describe and provides specific, controlled information which can be used subsequently.

4.2 Web directories for content qualification

As one of the tools for searching content and services on the Web, a directory provides the user with a hierarchical classification of sites grouped under thematic categories. Unlike search engines, Web directories give a universal description of reference sites, and provide users with a commented classification of them and organises them into categories and sub-categories.

We developed a program to collect, for a given Web directory, information on its structure (category and redirection structure) and the sites that it indexes (URL, title, description).We already described and compared six Web directories in [2]. The aim here is to use the textual description of the site or the page indexed in the directory and its position in the categories and sub-categories to characterise its content thematically and functionally. This method of content characterisation has several advantages: 1) there is no need to crawl the pages visited by users; 2) Web directory proposes a kind of categorization of the "Internet World" adapted to this particular object; 3) site and page descriptions are verified manually by the directory indexers. Conversely, this approach has certain disadvantages, the most important of which is that indexing is generally made at site-level and not at page-level.

When applying this method with two French-speaking directories observed in 2002 (Voila and Open Directory France), the coverage rates of our 2002 traffic data are about 27% for each. Considering their specificities, using more Web directories should lead to an overall coverage of 45% of the 27 million visited pages.

5. SESSION-LEVEL ENRICHMENT

After enriching data at the URL level, we can now aggregate these descriptions at the session level, as well as calculate session-scale specific indicators. First, it is necessary to choose the proper scale of analysis in relation to the type of information we have: page, site or even service (from *CatService*) level. We estimate that the site level, or the service level when provided, is preferable to the page level, which should be used carefully. Secondly, due to important differences in page conceptions, duration is a better key than page count for measuring user interest for a visited site.

Afterwards, in order to deal with the temporal and dynamic dimension of browsing, we have developed simple and robust statistical indicators representing the "shape" and the rhythm of a session: whether it is linear (each page/site is seen only once) or not; the number and length of detours; the importance of these detours in terms of session temporality; distinguish and quantify the concentration of revisits on hubs in the session. We also wanted qualitative information on the way that pages are revisited, and developed a specific algorithm capable of identifying "*Back button"* sequences and isolating them from the rest of the session.

6. CONCLUSION

Our platform was successfully tested on user-centric traffic data from the SensNet panels. Both Web directories and *CatService* proved to be efficient for describing Web contents. With this combined use of *CatService* and Web directories, the rates of session coverage by the descriptions are greatly improved: overall, 48% of the observed traffic is described in terms of duration, and 53% of the sessions are described for more than half of their duration. This platform was used to carry out a large number of usage studies in the context of the SensNet project. The precision of the content descriptions as well as the scalability of the *CatService* module allowed us to perform fine-grained focuses on particular kinds of sites or services: use of Web search engines [1], use of Digital Libraries [3]. Beside these specific studies, we combined content descriptions with topological and temporal indicators in a global approach of navigation, and showed off the strong link between page content, browsing dynamics and users' personal territories on the Web [4].

7. REFERENCES

[1] Assadi, H. and Beaudouin, V. Comment utilise-t-on les moteurs de recherche sur Internet ? Réseaux, 20 (116). 171-198.

[2] Assadi, H. and Beauvisage, T., A comparative study of six French-language Web directories. in ISKO 2002, (Granada, Spain, 2002), 271-278.

[3] Assadi, H., Beauvisage, T., Lupovici, C. and Cloarec, T., Users and uses of online digital libraries in France. in Research and Advanced Technology for Digital Libraries. 7th European Conference on Digital Libraries (ECDL 2003), (Trondheim, Norway, 2003), Springer.

[4] Beauvisage, T. Sémantique des parcours des utilisateurs sur le Web. PhD in Sciences du Langage, University of Paris X, Nanterre, 2004.

[5] Catledge, L.D. and Pitkow, J.E. Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27 (6). 1065-1073.

[6] Cooley, R., Mobasher, B. and Srivastava, J. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1 (1).

[7] Cunha, C., Bestavros, A. and Crovella, M.E. Characteristics of WWW Client-based Traces, Computer Science Department, Boston University, 1995.