Self-Tracking on the Web: Why and How

Mathieu d'Aquin, Matthew Rowe and Enrico Motta
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
{m.daquin,m.c.rowe,e.motta}@open.ac.uk

Social, professional or commercial interactions on the Web rely extensively on the exchange of private, personal information. This is already the case in the offline world where disclosing certain personal information is necessary to enable engagement with other people and organisations. However, on the Web, the circulation of such information is happening in an un-restrained, fragmented and distributed environment, making it difficult for individuals to monitor and control what is being exposed and shared about them. In other words, while personal information, interests and habits are being tracked by a large number of websites and organisations through various mechanisms and for various purposes, individual Web users are mostly unaware of the type of information they expose and that is circulated about them on the Web.

In this position paper, we argue for the need for better consideration of the activity of self-tracking - i.e., the activity of monitoring and analysing one's own behaviour regarding personal information exchange and the consequences of such behaviour on their exposure, privacy and reputation. Indeed, recently there have been growing concerns regarding the way personal information is handled by the organisations collecting it, and how such information could be used to the disadvantage of Web users. Amongst the most cited issues are identity theft, lateral surveillance and data aggregation to the benefit of commercial companies or for malevolent activities. However, as our preliminary experiments have shown [d'Aquin et al., 2010a], the inherent complexity and fragmentation of the flow of personal information on the Web makes it impossible for an individual Web user to monitor, make sense of and act on his/her own exposure without appropriate technological support. In contrast with such complexity, the tools currently available to Web users are extremely limited. More and more users would simply use popular Web search engines to check websites where their name appears, however with all the noise and ambiguities that such a method introduces [Madden and Smith, 2010] the effectiveness and success of such an approach is limited.

The requirement to achieve effective self-tracking appears with respect to such issues, in an environment as complex as the Web. It can be seen as a specific approach to lifelogging (called Web lifelogging in [d'Aquin et al., 2010a]) focusing on Web interactions, with the purpose of providing sufficient data to achieve appropriate levels of personal information management [Jones and Teevan, 2007], personal reputation management, and of course, privacy.

While appearing as such a crucial need, support for self-tracking on the Web has remained mostly unexplored, apart from isolated initiatives and tools focusing on specific issues. Here, we review such initiatives and tools with the aim to identify a path towards a more principled and comprehensive approach to self-tracking. We distinguish two major trends in existing work: tracking one's own behaviour in terms of Web interactions and exchange of personal information, and tracking the appearance of one's personal information on the Web.

Tracking one's own Web interactions, traffic, behaviour

Research, as well as many commercial developments, have until now mostly been dedicated to logging user visits to websites, in order to provide valuable information to website owners in the form of patterns of interactions. However, tools such as Google Web History[1] can be

[1]www.google.com/psearch

used to record different aspects of Web activities, as long as they are done in the scope of what can be perceived by Google systems. Such an approach provides an interesting starting point to collecting information regarding one's own behaviour online, but has obvious limitations, including the lack of comprehensiveness and control over what is being collected, as well as the need to go through a third party (Google).

The perceived gap in the ability of users to take ownership of their own Web activity data has led to the emergence of the notion of attention data[2], with tools such as the Attention Recorder[3] developed explicitly to provide the user with ways to track their Web activity, as carried out through a browser. The idea here is that the user can claim back their own activity data, so that they can be shared and traded in their own terms. Technically, tools such as the Attention Recorder still need to gain maturity, to be able to cover the wide variety of sources of activity (attention) data on the Web, and to provide appropriate support for the user to truly exploit the collected information.

In [d'Aquin et al, 2010a] we experimented with the idea of a complete, unrestricted 'self-monitoring' of personal, online activities, in a process comparable to the idea of *lifelogging* [O'Hara et al, 2009]. Even in relatively small settings, such an approach provides rich data about the user's behaviour [d'Aquin et al., 2010a], using a "local Web proxy" to obtain Giga Bytes of information about a single user's Web activities within the scope a 2.5 months. Specific analyses of the data collected revealed promising potential for such an approach. Simple geographical mappings of the requests from the user shown expected patterns, with most of the activities concentrating in Europe and North America, but also helped identifying anomalies (e.g., a small number of requests to Nigeria) that could be explored further based on the collected data. Looking at other indicators, such as the number of requests to different websites, the quantity of information transferred to these sites, and the user agents used in these transactions also demonstrated the extent to which activities and exchanges on the Web are "implicit", i.e., realized without being explicitly triggered by the user. More sophisticated analyses based on the keywords used to query search engines showed how such simple information can be used to build a profile of the interests of the user, according to a particular view which might not be the one he or she is prepared to expose. There is indeed a generalized discrepancy between the user's view of his/her own behaviour on the Web, and the reality of this behaviour as it can be perceived through self-tracking. To illustrate this point, in [d'Aquin et al., 2010b], we devised a model of the observed trust in websites and criticality of pieces of personal information, which is derived from the traces of activities collected for an individual user. The idea is that, through exposing users to such an abstract view of their own behaviour online, they can make emerge such discrepancies, leading to a better understanding and an improved awareness of the potential consequences of exposing personal information.

The idea of "logging" one's own Web activities is still in an early stage and the potential for analysis of such an approach remains mostly unexplored. In other terms, Web lifelogging faces similar challenges to other forms of lifelogging, including the need for mechanisms to abstract and interpret the obtained low-level raw data into something exploitable by the user [d'Aquin, 2010].

Tracking one's references on the Web

Besides tracking one's own behaviour, a key to self-tracking is the ability to monitor what information about an individual has been made visible on the Web, possibly without the user's consent. Web presence is an important aspect of business and reputation for the majority of Web users. Inflammatory content or misleading information can have dire consequences for the individual that it describes, for instance, [Andrejevic, 2005] cites

[2] see e.g., http://majestic.typepad.com/seth/2005/10/atx_the_attenti.html

[3] https://addons.mozilla.org/en-US/firefox/addon/3569/

examples of employers 'vetting' prospective employees by searching the Web for information about them. The recent Javelin report[4] describes the 2010 identify fraud statistics collected from US companies, showing an overall reduction in the number of cases, while the mean economic cost of such cases has risen – indicating a move towards targeting selective individuals. Individual web users must be informed where there personal information resides on the Web, so that the correct action may then be taken – i.e. applying for the information to be removed if it has been placed there without consent, or altering the visibility settings of the profile if the user has intentionally placed it there.

The sheer scale of the Web however makes manually finding web references largely infeasible. Automatic methods and third party services therefore provide a viable solution to overcoming such tasks. Identifying web citations is a single-person disambiguation task: given a collection of Web pages, all of which contain a specific person's name, the goal is to disambiguate those pages which refer to the individual of interest. Our experience [Rowe and Ciravegna, 2010] shows that an efficient approach is to use a combination of supervised classification models with a semi-supervised framework. A common issue when applying such methods is obtaining initial seed data to start the identification process. For instance, we may only know a few web references for the individual, the information from which we can use as seed data describing the person. Using such a framework, therefore, allows information to be learnt in an on-going process as more web references are found and the information within those web references put to use.

The extraction of such information also poses a problem. The messiness of information provided on the Web, given the heterogeneous nature of HTML and the lack of conformance to web standards, makes it hard for machines to parse web pages for personal information. Techniques are therefore required that can effectively extract personal information from the Web at high-levels of accuracy. Furthermore, as mentioned previously, information published on the Web about an individual may damage the person's reputation if it is negative or describes the individual in a bad way. Sentiment analysis techniques are therefore required which can assess the sentiment, or feeling, towards the person in the web page, enabling reputation assessment in an automated fashion at a large-scale.

Several companies have tackled the above issues, for example SentiMetrics[5] use social media sources to calculate the sentiment towards a given person based on available information, and Trackur[6] and Visible Technologies[7] also monitor social media sites for references to a person. Garlik's Data Patrol[8] service assesses the risk of an individual to identity theft, based on the presence of their sensitive information on the Web. Identity Guard[9] provides a service that monitors a person's information distributed across the Web, and alerts the individual when the exposure of his/her information could have a detrimental effect.

While such existing services tackle the individual aspects of web exposure, a single unifying approach is currently lacking that informs the web user where there personal information resides on the Web, the sentiment that such references have, and ultimately how the visibility of such information could effect the person. Therefore a core, unsolved challenge is to integrate and relate all these different pieces of information, to understand and interpret them in a context which takes into account the user's identity, activities and own perception of his or her exposure.

---

[4] http://www.idsafety.net/report.php

[5] http://www.sentimentmetrics.com/

[6] http://www.trackur.com/

[7] http://www.visibletechnologies.com/

[8] http://www.garlik.com/dpindividuals.php

[9] http://www.identityguard.com

Conclusion

More and more personal information is being shared, exchanged and exposed by Web users everyday, mostly without their consent and awareness. A lot of efforts and attention is currently being given to the way online organizations might track this information, to their own benefit, and potentially, to the detriment of the users. Here, we discussed initial tools and techniques towards taking the inverse perspective: helping Web users tracking and monitoring their own personal information online, to their own benefit.

As our initial experiments have shown, achieving such a process of self-tracking can be very revealing to Web users, helping them reaching a better awareness of their own online behaviour, and a better understanding of the possible consequences of such behaviour on the exposure of their personal information. Such an approach appears to be crucially needed as the Web evolves to both a global information marketplace, and a major medium for all sorts of social interactions online. However, the tools and technologies currently available to carry out self-tracking on the Web are inadequate, to the point that many Web users would resort to using a Web search engine to check where their name appears [Madden and Smith, 2010].

We therefore argue that a more principled and comprehensive study of the activity of self-tracking on the Web and of the technological requirements for such an activity to take place should be conducted. This requires for both the social and conceptual models of the way personal information is exchanged on the Web to be related to the technological protocols that are used as mediums for instantiating these models. From a more concrete point of view, we believe that a new set of tools are to be created that will support users in monitoring their own activity on the Web, tracking the appearance of their personal information online, and interpreting this information in terms of behaviour, reputation and privacy risks. A positive effect of the availability of such tools is not only to provide individuals with better control over the exposure of their information, but also to support a generic understanding of the global mechanisms underlying such circulation of personal information on the Web.

References

[Andrejevic, 2005] M. Andrejevic (2005) The work of watching one another: Lateral surveillance, risk and governance. Surveillance and Society, 2 (4):479–497, 2005.

[d'Aquin, 2010] M. d'Aquin, (2010) Making Sense of Users' Web Activity, Personal Semantic Data, PSD (keynote) at EKAW 2010.

[d'Aquin et al, 2010a] d'Aquin, M., Elahi, S. and Motta, E. (2010) Personal Monitoring of Web Information Exchange: Towards Web Lifelogging, Poster at Web Science 2010 Proceedings of the WebSci10: Extending the Frontiers of Society On-Line

[d'Aquin et al, 2010b] d'Aquin, M., Elahi, S. and Motta, E. (2010) Semantic Monitoring of Personal Web Activity to Support the Management of Trust and Privacy, Workshop: SPOT 2010 - 2nd Workshop on Trust and Privacy on the Social and Semantic Web at ESWC 2010

[Jones and Teevan, 2007] W. Jones and J. Teevan (editors) (2007), Personal Information Management, University of Washington Press

[Maden and Smith, 2010] M. Madden, A. Smith (2010), Reputation Management and Social Media, Report from the PewResearchCenter (http://www.pewinternet.org/Reports/2010/Reputation-Management.aspx)

[O'Hara et al, 2009] K. O'Hara, M. Tuffield and N. Shadbolt (2009) Lifelogging: Privacy and Empowerment with Memories for Life. Identity in the Information Society, 1 (2).

[Rowe & Ciravegna, 2010] M. Rowe and F. Ciravegna, (2010) Disambiguating Identity Web References using Web 2.0 Data and Semantics. The Journal of Web Semantics.