

Making Sense of Users' Web Activity

Mathieu d'Aquin

Knowledge Media Institute, The Open University, Milton Keynes, UK
{m.daquin}@open.ac.uk

Personal information management (PIM), as described by [1], is “*the practice and study of the activities people perform to acquire, organise, maintain, retrieve, use, and control distribution of information items*”. More and more services rely on the Web to communicate with their users. The way users can control the distribution of personal information exchanged daily through various Web channels therefore appears as a crucial task for PIM. However, while the definition above clearly covers such activities, PIM has traditionally been focusing more on the aspects of supporting information organisation and integration for the purpose retrieval. Indeed, the types of personal information mentioned in [1] include elements such as “*information about a person but kept by and under the control of others*”, but ignore one of the most difficult type of information to manage: *information about a person which is being shared and exposed to others*.

The related issues not only concern the ways to monitor, store and retrieve this specific type of information, but also the ways for users to make sense of the huge amounts of information they are exchanging on the Web, knowingly or unknowingly. Indeed, as a first building block in this area, we developed a tool dedicated to tracking the activity of an individual user on the Web. In practice, this tool takes the form of a ‘local proxy’ intercepting and storing (using Semantic Web standards) the HTTP traffic on the user’s computer. At a higher level, we can see this tool as a ‘Web Liffellogger’, dedicated to the indiscriminating collection of information concerning the user’s online activity. While relatively basic in principle, experimenting with this tool over a period of time generates huge amounts of data (100 Million Triples for a single user in 2.5 months) which, when studied, allows us to unveil interesting, and sometimes surprising aspects of the users Web life.

The use of semantic technologies offers the right level of flexibility for the management of such large, heterogeneous data, but more importantly, provides us with the data integration and modelling approaches necessary to making sense of the data. For example, mapping the collected semantic logs with a representation of the user profile allows us to construct models of the perceived trust the user gives to various websites regarding the handling of his/her personal information, and of the sensitivity of this information. Going a step further, by applying different ontologies over the data, and linking it to the Web of Data, we can build different perspectives on the traces of Web activity produced by the user, providing as many “interpretations” of the user’s interaction with the Web, in addition to tools supporting him/her in managing this interaction.

1. William Jones and Jaime Teevan (editors), *Personal Information Management*, University of Washington Press, 2007