I remember I have seen this: how can I re-find it?

Samur Araújo

Delft University of Technology, PO Box 5031, 2600 GA Delft, the Netherlands
s.f.cardosodearaujo@tudelft.nl

Abstract. Everybody experiences it: our personal information from mails, documents, chats and browsing is scattered over many applications, formats, and devices, which makes it difficult to access it again some time after it has been seen for the first time. The challenge of the Memorizer project is to allow users to re-find a piece of data that they have seen before but for which they cannot remember where it is. This objective can be achieved by building a Personal Web of Data, a semantic layer that describes, aggregates, interlinks and stores metadata about information that users have accessed before. Later on, this web of data can then be exploited using smart semantic browsing. In this PhD symposium paper we consider Memorizer's grand challenge in re-finding and the approach to achieve this in the period ahead.

Keywords: semantic browsing, semantic web in use, information extraction, memory, user history, linked data, RDF.

1. Introduction

"I remember I have seen this: where is it?" This question is asked everyday by millions of people that work in a digital information environment [1, 2]. The core challenge for solving this problem is to allow users to find back information that they have seen before, using any mental clue (represented through meta-information or metadata [3]) they can remember, independent of the application, format or device where the information was seen before. Our hypothesis is that bringing the representation of the users' personal data closer to the users' mental model about that data, or their memories, we can better support the re-finding tasks. Semantic Web technology can play a crucial role in this challenge, since it provides a complete framework to weave the users' personal data into a web of data that, in level of granularity, is closer to the users' memory.

Many important related issues have been tackled in the Personal Information Management (PIM) research area [4, 5, 6], e.g. finding and re-finding of information, representing and unifying of personal information, keeping and organizing of personal information, privacy, security, etc. Nevertheless, the role of semantics and in particular the elicitation of metadata in the context of re-finding has been dealt with in a modest manner in that area. Firstly, because semantic technologies are still reaching maturity and there is no reference architecture about how to apply natural language processing, information extraction, Resource Description Framework (RDF) model

[7], linked data, ontology matching, etc. for that concern of metadata elicitation for re-finding. Another reason is the fact that the PIM research area is still learning how people collect, store, manage, and specially re-find information.

In spite of several projects that demonstrate the benefits of bringing together semantic technology and PIM, such as Nepomuk [8], Semex [9], IMemex [10] and Haystack [11], these works were not particularly focused on *re*-finding and many aspects remain to be addressed, such as information fragmentation [12], addressability, mobility of data, and also privacy, provenance and trust [13, 14, 15] that we will not concern in this paper for the matter of space.

In this PhD work, we are concerned with the use of Semantic Web in the context of re-finding digital information. In particular, we focus on the problem of data access and show that the elicitation, combination and interlinking of metadata using the semantic approach can be applied to enrich and optimize re-finding of information.

For instance, suppose that Ana is designing a web site together with Daniel. Daniel finds a reference that may be interesting to Ana: the Scriptaculous API, a Javascript API for building animations. So, he sends an e-mail recommending her to have a look on the Scriptaculous' web site. As Ana has already decided which technology to use in the project she does not visit the web site. One year later, Ana is looking for a Javascript API for building pop-up menus on web pages. She has visited many pages but she cannot find any solutions suitable to her needs. Fortunately, she remembers that Daniel has mentioned an API (the Scriptaculous) that can be useful, however, she does not remember the name of the API, or the URI or when Daniel gave her such information. In this scenario, the system can use the user mental clues during the re-finding task by eliciting and connecting information between the user history and email messages. For instance, Ana can search for messages sent by Daniel, but as the system knows that she has just visited few pages concerning Javascript APIs, the system can filter messages sent by Daniel that also are related to Javascript concept. As result, Ana will be able to find the Scriptaculous URI by providing only her mental clues that she is able to remember, instead of wasting her time by browsing the entire mailbox or performing many queries on search engines.

2. State of the Art

Due to the nature of this PhD symposium paper, an exhaustive overview on related work cannot be given here. Instead, we focus on the identification of some specific problems that allow us to help clarify the need for a semantics-based approach.

Information Fragmentation

The process of re-finding personal data is closely related to the process of how the personal data was collected and stored. Unfortunately, the tools that users have today, lead them to spread their data among many applications and formats, therefore decreasing and/or breaking the data connectivity. For instance, once a user saves an email attachment to her file system, the connection between the file and message is lost. Consequently, this demands a high mental load for re-finding the knowledge

about who sent that document. This lack of data connectivity is known in the literature as information fragmentation [12].

To solve this problem some authors [8, 9] have proposed to build an integration layer that describes the information in a uniform language. The Semantic Web (SW) [16] can play an important role in this endeavor, since the users' mental model of their digital artifacts can be expressed using the Resource Description Framework (RDF) model, that it is flexible enough to change, grow or expand the users' mental model representation as it evolves. The benefits of using these semantic technologies for building a uniform representation layer was shown in Nepomuk [8], Semex [9], IMemex [10] and Haystack [11]. Although, it has not been proven how good these systems perform the re-finding task, clearly there is space for improvement. They do not detect the user context that we argue is crucial for providing a better re-finding system. If the user is looking for something for the second time, we can infer that what she is looking for is based on her current context, since such context was seen at least once before, when the information was firstly accessed. Those semantic desktops do not exploit such correlation that can be used, at least, to reduce the amount objects presented for users during the re-finding task, consequently reducing the user mental load on such process.

Information Extraction

An important part of the problem concerns the techniques to elicit relevant metadata about the information that it is already there, e.g. file system, email, web history, etc. Information extraction in connection to the Semantic Web can be used to help in structuring the users' data that is likely wrongly (less optimally) structured for the purpose of re-finding, consequently allowing to form an integrated and universal medium of information that (expectedly) improves the possibility to re-find data [18]. Nepomuk, Semex, Memex, IMemex and Haystack use information extraction to build a semantic layer of metadata: however, they do not focus on combining and integrating the metadata to better support users re-find information.

3. Approach and Methodology

Little attention has being given until now to the process of *re*-finding information. The previous section indicates that there is still a lot of work to do regarding this challenge. In this PhD research we have started work on an approach centered around the notion of elicitation of metadata about information that user accesses, organizing it and exploiting it, in order to support re-finding tasks. In particular, we are concerned with the following research questions.

1. How do we integrate meta-data about the personal data for re-finding?
2. How do we query the integrated meta-data for supporting re-finding tasks?

To outline a realistic Ph.D. proposal, we will focus on re-finding of email and web history. Our first approach for that problem is to elicit meta-data about the user's web history and email archive. For that purpose, we are going to use available tools on the

literature that can be used for extracting information from the user's email and web history. Also, we are going to express and consolidate this meta-data in a uniform semantic layer, expressed in RDF model, which serves as a first step towards organizing the entities for re-finding. We are going to investigate the process of integrating data, and how this meta-data can be automatically interconnected. We will also investigate how to detect the user context based on her web navigation, and how to integrate such information with other elicited meta-data during the re-finding process.

All this information will be expressed in a novel ontology-independent approach. Our intention is to evaluate an approach where different terminologies can co-exist but where the system exposes to the user only a joint view of concepts, properties and classes that are similar but they were expressed in different terminologies. With this approach, the system can learn the user's mental model and adapt its terminology to it, instead of forcing the user to adapt to the system terminology.

Since there might be information that can be a crucial mental clue for re-finding but that can not be directly obtained from the user's personal data information, we are going to use external sources of data as WordNet[1], DBPedia[2] and other datasets on the Linked Data to enrich the semantic layer.

Once we have organized the semantic layer, we will expose it to users using different exploration mechanisms, such as browsing, facet navigation and keyword-search, in order to identify the best retrieval mechanism to perform and support re-finding tasks. We will evaluate the precision and recall of each retrieval mechanism.

4. Conclusion

We motivated our work with general considerations about the use of the semantic web in the process of personal information management. While there are many open problems, we focused on a specific family of interrelated problem that are centered around the notion of *re*-finding information using metadata about the information that they have been seen before. As a result, we expect to bring a novel architecture to represent the user's personal data that it is closer to the mental model that the user has about her environment, therefore improving how the information can be re-found. Nevertheless, we are still in early stage of development; we expect soon the delivery of the first tools that will be part of the whole re-finding system called Memorizer.

References

1. Tauscher L., Greenberg S., How people revisit Web pages: empirical findings and implications for the design of history systems, International Journal of Human-Computer Studies, v.47 n.1, p.97-137, July, 1997

[1] http://wordnet.princeton.edu/wordnet/

[2] http://dbpedia.org/About

2. Teevan J., Adar E., Jones R., Potts M., History repeats itself: repeat queries in Yahoo's logs, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA
3. Candan K. S., Liu H., Suvarna H., Resource description framework: metadata and its applications, ACM SIGKDD Explorations Newsletter, v.3 n.1, July, 2001
4. Al-Fedaghi, S. and Ahmad, M., Personal information modeling in semantic Web, The Asian Semantic Web Conference (ASWC), Beijing, China, 3-7 September, pp.668-681, 2006.
5. Sauermann L., Elst van L., Dengel A., PIMO - a Framework for Representing Personal Information Models. 9th International Conference on Knowledge Management and Knowledge Technologies 2-4 September 2009, Graz, Austria
6. Norrie, M. C., PIM Meets Web 2.0. In Proceedings of the 27th International Conference on Conceptual Modeling (ER 2008), Barcelona, Spain, October, 2008.
7. RDF Primer - http://www.w3.org/TR/rdf-primer/
8. Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G., and Gudjonsdottir, R., The NEPOMUK Project - On the way to the social semantic desktop. In Proceedings of the International Conference on Semantic Technologies (I- Semantics). 201-211. 2007
9. Cai Y., Dong X. L., Halevy A., Liu J. M., Madhavan J., Personal information management with SEMEX, Proceedings of the 2005 ACM SIGMOD international conference on Management of data, June 14-16, 2005, Baltimore, Maryland
10. Dittrich J.P , Salles M.A.V , Kossmann D., Blunschi L., iMeMex: escapes from the personal information jungle, Proceedings of the 31st international conference on Very large data bases, August 30-September 02, 2005, Trondheim, Norway
11. Karger, David R. and Bakshi, Karun and Huynh, David and Quan, Dennis and Sinha, Vineet., Haystack: A Customizable General-Purpose Information Management Tool for End Users of Semistructured Data. Proceedings of the 2003 CIDR Conference. 2005
12. Tungara M., Pyla P., Sampat M., Perez-Quinones M., Defragmenting Information using the Syncables Framework, In: SIGIR Workshop on Personal Information Management, 2006
13. P. Nasirifard, M. Hausenblas, and S. Decker., Privacy Concerns of FOAF-Based Linked Data, In Trust and Privacy on the Social and Semantic Web Workshop (SPOT 09) at ESWC09, Heraklion, Greece, 2009.
14. H. Halpin., Provenance: The missing component of the semantic Web for privacy and trust, In Trust and Privacy on the Social and Semantic Web (SPOT2009), workshop of ESWC 2009, Heraklion, Crete, Greece, June 2009.
15. Jeremy J. Carroll, Christian Bizer, Pat Hayes, Patrick Stickler, Named graphs, provenance and trust, Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan
16. Nivas, I., Semantic Web Architecture, 6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007). Pages 77-78. November 2007.
17. Elsweiler, D., Baillie, M., and Ruthven, I., Exploring Memory in Email Re-finding. ACM TOIS CFP special issue on Keeping, Re-finding, and Sharing Personal Information. vol. 26 (4), pp. 1-36. 2008.
18. Bernstein M., Kleek M. V., Karger D., Schraefel M. C., Information scraps: How and why information eludes our personal information management tools, ACM Transact September, 2008