Collecting and Analysing Personal Information Management Data

Charlie Abela[1], Chris Staff[1], and Siegfried Handschuh[2]

[1] Department of Intelligent Computer Systems,
University of Malta, Malta
{charlie.abela,chris.staff}@um.edu.mt
[2] Department of Computer Science and Mathematics,
University of Passau, Bavaria, Germany
{siegfried.handschuh}@deri.org

Abstract. Personal Information Management (PIM) research has investigated the information trail generated by an individual while performing some information-seeking task on their desktop, with the aim of improving PIM tool-support. Nevertheless, due to the personal nature of the data, this is rarely released for reuse. Furthermore, there exists no tool that allows a PIM researcher to investigate how PIM related data evolves over time nor one that allows for the results of applying different approaches over such data to be analysed. In this paper, we present the Personal Information Management Analytix framework (PiMx) that leverages upon a graph-analytics approach for the analysis and visualisation of evolving activity-data generated by individuals performing tasks on their desktops. We further describe a data collection methodology that opens up the data for reuse and briefly discuss how PiMx is used to analyse such a collection.

Key words: Graph analytics, Personal Information Management, Task identification

1 Introduction

When we perform some information-seeking task we tend to spend a considerable amount of time looking back, establishing past references and remembering [4]. To find and re-find information items, we tend to rely on our organisational skills and the support of search, bookmarking and history tools [3]. However, most of these tools tend to consider the user's information-seeking activities as unrelated events, unlike the way we actually organise things, which is usually in terms of directories (on our desktop) and tasks (conceptually) [5].

In our research we are motivated by the need to better understand how these activities evolve over time and the extent to which it is possible to automatically organise them in terms of tasks. In this paper we present *PiMx*, a **P**ersonal **i**nformation **M**anagement analyti**x** framework that we implemented to support us in our investigation. *PiMx* enables us to simulate the incremental evolution of PIM data and to exploit graph-analytics to analyse and visualise the user's information-seeking process. It is also possible to apply different algorithmic approaches and analyse their performance in addressing the task-identification problem. To the best of our knowledge, no such tool is available.

It is difficult to find suitable PIM datasets freely available for reuse and evaluation. We therefore performed a controlled experiment to collect our own dataset. We briefly elaborate on the adopted methodology and describe how we use PiMx to analyse and compare a number of approaches scoped at automatically identifying task-clusters from the data.

In the rest of paper we provide some related work in Sec. 2 which is followed by a description of the controlled experiment we performed and the data that was collected. In Sec. 4 we give an overview of the PiMx framework used to analyse the collected data and conclude with some future work.

2 Related Work

The task modelling ontology proposed by [7] provides task-related information support for knowledge workers and links the user's task activities with her personal information context. The user's desktop activity context was also modelled by [8] as an OWL-DL ontology and used to enhance the performance of task detection algorithms. A similar context model was proposed by [9] which defines events and contextual elements relevant to a knowledge worker and so needs to also deal with projects and collaborative work. We have adopted this latter model for our activity data.

Evaluating PIM tool-support is inherently difficult, in particular because of the lack of readily available datasets [3]. One available dataset is provided by the Web History Repository project[3]. Participants in the project can voluntarily relinquish their Web browsing history which is anonymised and remotely sent to a server via a dedicated Firefox[4] plug-in. Each user's history is uniquely identified by a global ID and the URLs accessed are encrypted.

Existing graph analysis tools, such as Visone[5] and Gephi[6], do not support the analysis, over time, of streams of users' activity nor is it possible to apply and compare different user-defined algorithms over the streams.

3 Collecting and Modeling Activity Data

Although the dataset from the Web History Repository is substantial, we were unable to use it as, for our research, we need to know, for each user, the sequence in which documents were accessed as well as the tasks the user was engaged in when accessing the documents. Thus, we conducted our own data collection

[3] http://webhistoryproject.blogspot.com/

[4] https://www.mozilla.org/en-US/firefox/desktop/

[5] http://visone.info/

[6] http://gephi.github.io/

experiment in a controlled environment during which we logged the browsing activity of 20 participants while performing three pre-defined tasks. The tasks were: providing specific information about the planning of a *vacation* in a specific country; answering questions related to the research area of *human computation*; and, providing information about any two upcoming *music events*.

For the experiment, we set up a cluster of machines in one of our laboratories. Each PC ran Windows OS with two activity-monitoring applications installed on them. The first application was a Firefox plug-in used to collect each participant's Web browsing activity. The second application monitored the file browsing activity (such as word processing documents) on their desktop. We cleaned the data and removed references that could lead to the identification of the participants. References to the accessed documents were however retained. In the future, the dataset can be made accessible and shared for research evaluation purposes, in line with the Web History repository's philosophy.

The logged data is represented in RDF and is based on the context model developed by [9]. It includes information about the type of event, such as whether it is a navigational or a tabbed event; the application that generated the event; the timestamp; the URI of the document accessed as a result of the event; and an excerpt of text from the window caption. Other information specific to particular events is also captured. This includes the URI and window caption of the page that was in focus before the event was triggered and information about files found on the user's desktop such as the file name and whether a document was edited or not. The example shown in Listing: 1.1 represents an instance of an *EnteredURL* event which is generated whenever the user manually enters the URL in the browser's address bar.

```
<http://test.org/actions/EnteredURL20140430T134829>
  a <http://test.org/vocabulary/actions/EnteredURL> ;
  actions:timeStamp "2014-04-30T13:48:29"^^xsd:dateTime ;
  actions:processName "firefox"^^rdfs:Literal ;
  actions:uri <https://www.google.com.mt/search?q=carl+bee+song> ;
  actions:docInfo "carl_bee_song"^^rdfs:Literal ;
  actions:fromURI <http://www.tvm.com.mt/news-isle-of-mtv-malta/> ;
  actions:fromPageTitle "Isle_of_MTV_Malta"^^rdfs:Literal .
```

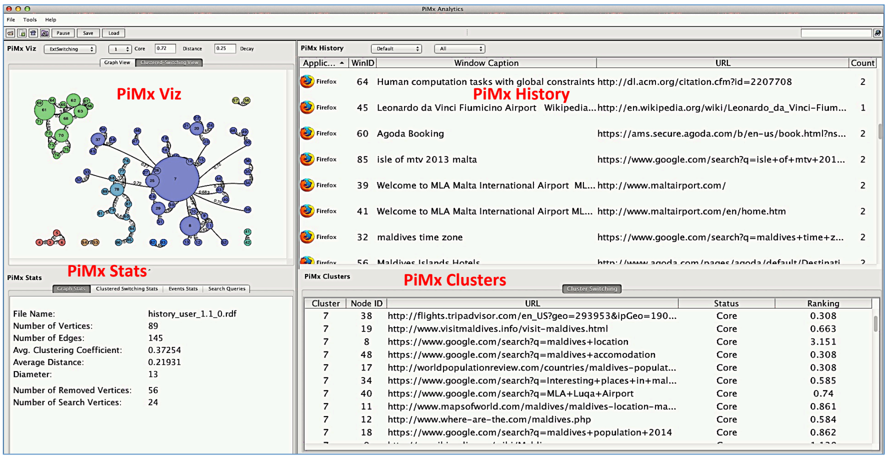
Listing 1.1: EnteredURL Event

4 PiMx: tool for analysing the data

We implemented the PiMx (**P**ersonal **i**nformation **M**anagement analyti**x**) framework to better analyse the collected data. This tool enables us to load a user's activity-log (collected during the experiment described in Sec. 3), simulate the user's behaviour by replaying the activity trail for that user and analyse the evolution of the task-clusters through different views. This process can be paused and resumed at any stage. PiMx uses the JUNG graph library[7] for the graph-analytics and Apache Jena[8] for modeling and querying the data.

[7] http://jung.sourceforge.net/

[8] https://jena.apache.org/



File Name: history_user_1.1_0.rdf
Number of Vertices: 89
Number of Edges: 145
Avg. Clustering Coefficient: 0.37254
Average Distance: 0.21931
Diameter: 13
Number of Removed Vertices: 56
Number of Search Vertices: 24

Applic...	WinID	Window Caption	URL	Count
Firefox	64	Human computation tasks with global constraints	http://dl.acm.org/citation.cfm?id=2207708	2
Firefox	45	Leonardo da Vinci Fiumicino Airport Wikipedia...	http://en.wikipedia.org/wiki/Leonardo_da_Vinci-Fium...	1
Firefox	60	Agoda Booking	https://ams.secure.agoda.com/b/en-us/book.html?ns...	2
Firefox	85	isle of mtv 2013 malta	https://www.google.com/search?q=isle+of+mtv+201...	2
Firefox	39	Welcome to MLA Malta International Airport ML...	http://www.maltairport.com/	2
Firefox	41	Welcome to MLA Malta International Airport ML...	http://www.maltairport.com/en/home.htm	2
Firefox	32	maldives time zone	https://www.google.com/search?q=maldives+time+z...	2
Firefox	56	Maldives Islands Hotels	http://www.agoda.com/pages/agoda/default/Destinati...	2

Cluster	Node ID	URL	Status	Ranking
7	38	http://flights.tripadvisor.com/en_US?geo=293953&ipGeo=190...	Core	0.308
7	19	http://www.visitmaldives.info/visit-maldives.html	Core	0.663
7	8	https://www.google.com/search?q=maldives+location	Core	3.151
7	48	https://www.google.com/search?q=maldives+accomodation	Core	0.308
7	17	http://worldpopulationreview.com/countries/maldives-populat...	Core	0.308
7	34	https://www.google.com/search?q=Interesting+places+in+mal...	Core	0.585
7	40	https://www.google.com/search?q=MLA+Luqa+Airport	Core	0.74
7	11	http://www.mapsofworld.com/maldives/maldives-location-ma...	Core	0.861
7	12	http://www.where-are-the.com/maldives.php	Core	0.584
7	18	https://www.google.com/search?q=maldives+population+2014	Core	0.862

Fig. 1: PiMx Interface

PiMx includes an interactive *PiMx-Viz* component which currently presents two visualisations (see Fig. 1). The first visualisation shows the complete, unadulterated activity-log as an undirected graph that evolves over time. The second displays the coloured task-clusters as they are incrementally created. Nodes are assigned a global unique ID and their size is computed in relation to a ranking value. Edges are weighted based on the number of switches between any two nodes. Further information about the nodes and edges, such as the node's degree and its URL, as well as the type of edge and timestamp of last access can be viewed by hovering over them. It is also possible to click on each node separately and visualise the induced subgraph generated by the nodes' neighbourhood.

The *PiMx-Stats* component (Fig. 1) shows different graph-related statistics, such as the number of vertices and edges in the graph and the clustering coefficient, average distance, and diameter. There is also information about the number of search and removed nodes. The former represent the pages associated with search queries while the latter refer to those documents that were closed by the participants after being accessed. This view also provides information about the type and number of occurrences of the events that were triggered.

It is possible to incrementally view a detailed history of all the accessed documents through the *PiMx-History* view. This information includes the application used to access a document, the window caption, the time of the last access, the URL and the number of times that a document was accessed. The researcher can apply filters over this data and view it based on different time-windows, such as by last hour, last 4 hours, today and yesterday, as well as by application or file-type. A search facility based on Jena-Fuseki's text query and a Lucene index allows for keyword search over the data.

The *PiMx-Clustering* component is specific to the goal we wanted to attain and provides information about the automatically generated task-clusters. Each cluster is assigned a unique ID and each document node within a cluster has associated with it a ranking value based on its importance within that cluster.

Through PiMx we are currently able to compare the suitability of different algorithmic approaches, in particular we applied the Bron-Kerbosch algorithm for finding maximal cliques [2], the community detection algorithm developed by [6] and our **iDeTaCt** density-based clustering algorithm, details of which can be found in [1].

5 Conclusion

In the near future we will be collecting more PIM data and making it available for reuse. We also plan to provide extensibility interfaces for PiMx so that other evolving data such as social-network and web-usage data, can be visualised, analysed and searched.

References

1. Abela, C., Staff, C. and Handschuh, S.: Automatic Task-Cluster Generation based on Document Switching and Revisitation. In Proceedings of the 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP'15. (2015)
2. Bron, C. and Kerbosch, J.: Algorithm 457: finding all cliques of an undirected graph. In Commun. ACM 16, no. 9: pp. 575-577. (1973)
3. Jones, W.P., Teevan, J.: Personal Information Management. ISBN 9780295987378, University of Washington Press (2007)
4. Mayer, M.: Web History Tools and Revisitation Support: A Survey of Existing Approaches and Directions. Found. Trends Human-Computer Interaction, volume 2, 3, pp. 173-278 (2009)
5. Morris, D., Ringel Morris, M., Venolia, G.: Searchbar: a search-centric web history for task resumption and information re-finding. In: 26th annual SIGCHI conference on Human factors in computing systems, CHI '08, pp. 1207-1216. ACM Press, New York, NY, USA (2008)
6. Newman, M. E. J. and Girvan, M.. "Finding and evaluating community structure in networks." Physical Review E 69 , no. 026113 (2004)
7. Ong, E., Riss, U., Grebner, O., and Du, Y.: Semantic Task Management Framework, in K. Tochtermann and H. Maurer, ed., 'I-KNOW '08 Proceedings of the 8th International Conference on Knowledge Management, Graz, Austria (2008)
8. Rath, A.S., Devaurs, D. and Lindstaedt, S.N.: UICO: an ontology-based user interaction context model for automatic task detection on the computer desktop. In Proceedings of the 1st Workshop on Context, Information and Ontologies (CIAO '09). ACM, New York, NY, USA (2009)
9. Schwarz, S.: A Context Model for Personal Knowledge Management. Applications. In Modeling and Retrieval of Context, volume 3946, Springer, Berlin Heidelberg (2006)