YourDigitalSelf: A Personal Digital Trace Integration Tool

Varvara Kalokyri
Department of Computer Science
Rutgers University
v.kalokyri@cs.rutgers.edu

Alexander Borgida
Department of Computer Science
Rutgers University
borgida@cs.rutgers.edu

Amélie Marian
Department of Computer Science
Rutgers University
amelie@cs.rutgers.edu

ABSTRACT

Personal information is typically fragmented across multiple, heterogeneous, distributed sources and saved as small, heterogeneous data objects, or traces. The *DigitalSelf* project at Rutgers University focuses on developing tools and techniques to manage (organize, search, summarize, make inferences on and personalize) such heterogeneous collections of personal digital traces. We propose to demonstrate *YourDigitalSelf*, a mobile phone-based personal information organization application developed as part of the *DigitalSelf* project. The demonstration will use a sample user data set to show how several disparate data traces can be integrated and combined to create personal narratives, or coherent episodes, of the user's activities. Conference attendees will be given the option to install *YourDigitalSelf* on their own devices to interact with their own data.

ACM Reference Format:
Varvara Kalokyri, Alexander Borgida, and Amélie Marian. 2018. *YourDigitalSelf*: A Personal Digital Trace Integration Tool. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital traces of our lives are now constantly produced by various connected devices, internet services and interactions. Our actions result in a multitude of data objects, or traces, kept in various locations in the cloud or on local devices: messaging and email, calendars, location checkins (e.g., Facebook Places, Foursquare/Swarm or GPS tracker), online reservations (e.g. Opentable, Ticketmaster), reviews (e.g, Tripadvisor, Yelp), purchase history (e.g. Amazon, credit card statements), financial transactions, web searches, etc. These traces reflect a chronicle of the user's life, keeping record of where the user went, who the user interacted with (online or in real-life), what the user did, and when.

These "personal digital traces" are different from traditional personal files; they are typically (but not always) smaller, heterogeneous, and accessible through a wide variety of different portals and interfaces, such as web forms, APIs or email notifications; or directly stored in files used by apps on our devices. Personal digital traces can be connected into coherent groupings or episodes, for the purpose of exploring, remembering and understanding past user actions. For instance, a trip will result in many individual digital traces: email confirmation of hotel and/or flights reservations, financial transactions, GPS locations, messages and social media mentions of the trip, and more.

Given the disparate and disconnected nature of these personal digital traces, there is a dire need for novel techniques and tools that aid us in understanding and organizing them. The *DigitalSelf* project at Rutgers University aims to address this need, developing methods that enable users to understand their personal digital data by integrating personal digital traces into a unified whole. Based on this, entities and events in our data are connected into narratives which correspond to a user's autobiographical memories. This will allow users to explore their personal data and help them remember their digital memories as well as allow researchers to perform in-depth cross-service analysis and studies of user behaviors on various forms of services.

In this demo, we present the *YourDigitalSelf* tool, an Android mobile device application that gathers and integrates personal digital traces into coherent groupings that share a common theme, task or goal. This app is used as the basis to retrieve personal digital data originating from various sources to be used to implement and evaluate the research of the *DigitalSelf* project, through user studies and surveys. It can also act as a standalone user interface to help users navigate through their own personal digital traces in order to make sense of their data and individual patterns as well as to provide them with narrative views of their digital memories.

2 ARCHITECTURE

The architecture of the *YourDigitalSelf* app is shown in Figure 1. The View component is responsible for all the information presentation (UI layout, colors, borders etc) as well as handling all the UI changes (UI logic). The Activity/Fragment component in combination with the Application Manager contain all the business logic of our application. Data is stored on the user device both as files and SQLite data tables. (We are investigating the use of a JSON native storage database, however preliminary tests have shown that converting the JSON information received from APIs to relational tables stored in an Android-based SQLite is more efficient.)

2.1 Personal Data Extraction and Integration

YourDigitalSelf gathers various types of personal digital traces pertaining to a user and considers various sets of sources and objects, including emails, social network interactions, application data, web search information, financial data, and files.

Since there is a wealth of personal data and new sources of data are continuously appearing, we have limited ourselves to a number of popular services, but new services can be added as needed. We built upon our experience creating a personal information extraction desktop tool [11] to build a scalable (in the number of sources), dynamically maintainable system. The data currently collected in

Conference'17, July 2017, Washington, DC, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

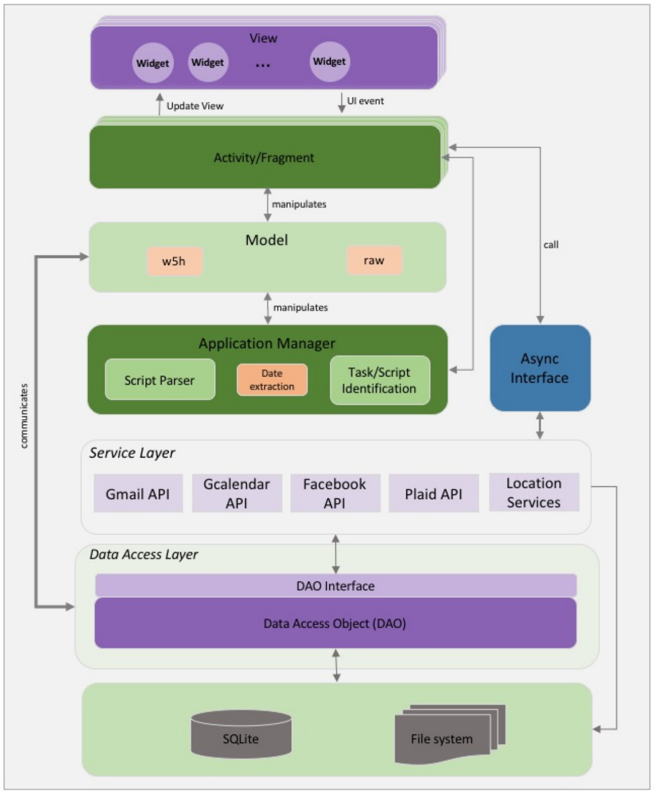


Figure 1: *YourDigitalSelf* architecture.

the application are from the following services: Gmail, Facebook, Instagram, Google Calendar, Plaid (for getting financial transactions), and GPS mobile data. It is also included the option of getting financial transactions through Google Drive, by giving users the ability to upload their financial transactions through .csv files given by various bank institutions.

One of the main challenges in integrating personal digital traces lies in the fragmentation of data and the heterogeneity of their data models. To integrate all these "document schemas", we have designed a unified and intuitive formal conceptual data model to link and represent both personal data and their corresponding coherent episodes [5].We base our model on observations in the Cognitive Psychology literature [1] that suggests that a natural way for people to remember past events is by any pertinent contextual information. Thus, personal digital traces can be modeled, and indexed following six dimensions that mirror the basic interrogative words: *what* (content), *who* (with whom, from whom, to whom,...), *where* (physical or logical, in the real-world and in the system), *when* (time and date, but also what was happening concurrently), *why* (sequence of data/events that are connected), *how* (application, author, environment) . We consider *how* to say the manner in which the information was recorded (e.g., device, or application), and *why* to be the task that involved the recording of the information (e.g., email and financial transaction related to the planning of a trip). We call this model the *w5h* model [4, 11].

In addition, a common challenge in personal data sets is that there are many same real word entities that appear under different names in the various services. To identify and connect contextual data, we have used and have adapted various existing data integration [2] and entity resolution and record linkage techniques [3, 6, 7] such as SERF [10]. This work takes place in the data model layer (Figure 1).

After the data is retrieved, mapped into our model, and entities are defined, the next step is to design an indexing structure that will facilitate the querying of the data. Toward this end, we use a mix of full-text and column indexing techniques provided by Android. The indexing takes place whenever the user decides to incorporate data in the app, either manually or by allowing the app to automatically get new data after a fixed period of time (e.g. every week/month).

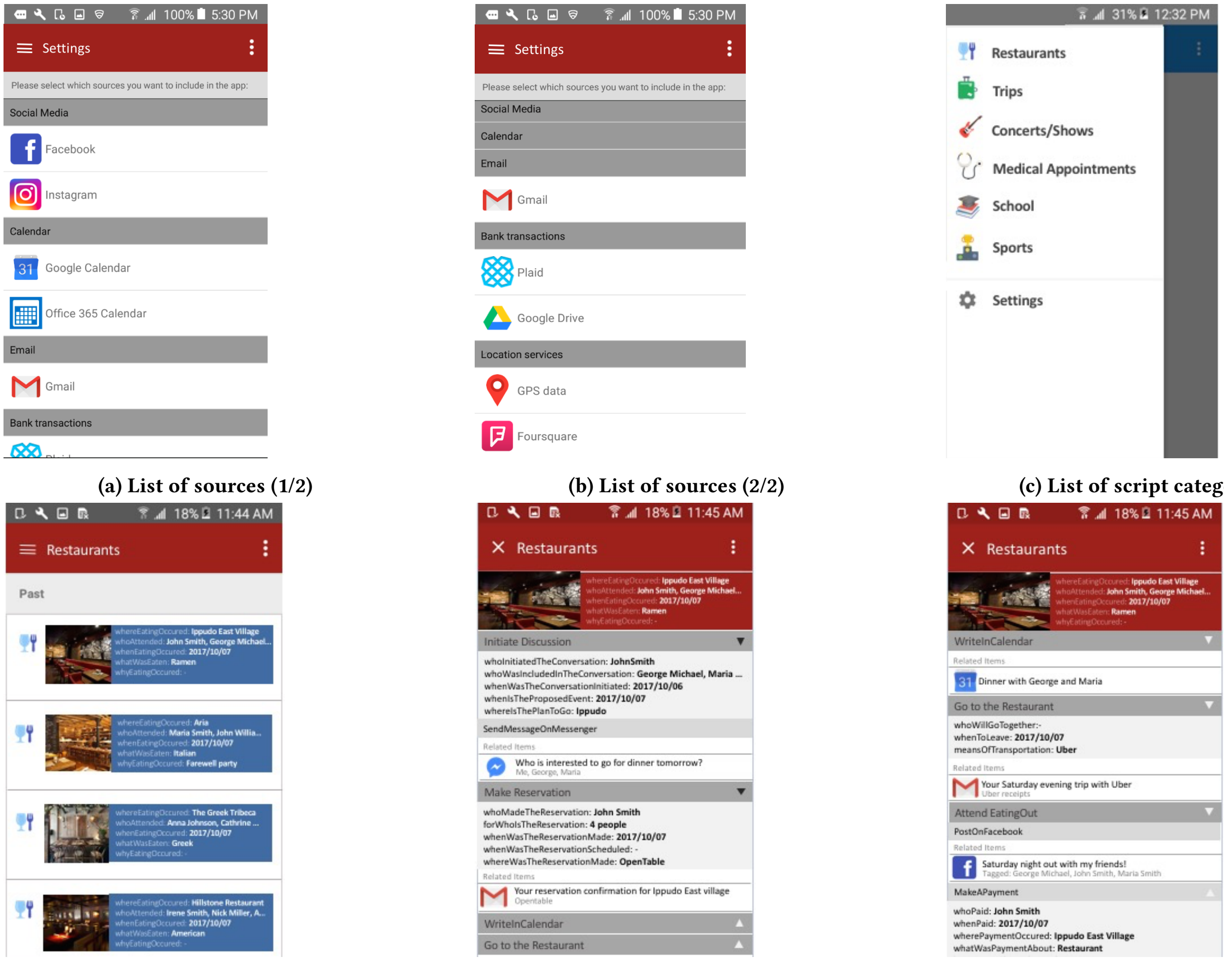
2.2 Event and Activity Recognition

Once the personal data traces have been integrated in our *w5h* model, a connection must be made between events (e.g., going out to eat), which are described by *scripts*/prototypical plans (see below), and personal digital traces (e.g., possible emails referring to this event, reservation, restaurant payment record). Essentially, each data trace presents *evidence* of various strengths for the occurrence of 0 or more atomic event instances, which make up the (sub)scripts (e.g., the same email may mention a restaurant and a theater play). Our goal is to group and connect related personal digital traces to candidate script instances to which they are likely to be associated.

The idea of scripts describing frequently occurring episodes is inspired by the work of Schank in AI [9], but has been considerably extended. Scripts have their own descriptive properties (e.g., who made the reservation, at what time, for what date), associated “parts” (subscripts) bottoming out at atomic actions, and expressions describing valid sequences of actions (e.g., making a reservation to the restaurant must precede paying for the dinner there, but discussing what is going to be eaten and the date can be carried out concurrently). We have applied the six interrogative dimensions described above scripts and their various components, thereby further integrating the information according to *w5h* axes [5]. We also provided a declarative language for expressing script descriptions, including the manner in which properties of a subscript (e.g., whose credit card was used to pay at the restaurant) are likely to be related to properties of the higher script (e.g., who went out out to eat), thus facilitating description, change and inference [5].

Most importantly, we developed an algorithm for constructing scripts instances based on the likelihood of each property [5]. For each sub-script we examine weak and strong evidence, as determined by the script description (e.g., a payment for a restaurant is very strong evidence for instantiating the "Eating out" script whereas a discussion about a restaurant is weak evidence), and the likelihood of each property, and merge this information to instantiate the top-level script). Note that an individual trace may instantiate several different sub-scripts, which in turn may be part of separate scripts. Our bottom-up approach uses document search for keywords that are selected in a principled manner for each script, based on vocabularies and ontologies.

Our system can be extended to cover many different scenarios/scripts due to the systematic and extensible approach of our script creation and instantiation. All the scripts definitions are



(a) List of sources (1/2)

(b) List of sources (2/2)

(c) List of script categories.

whereEatingOccured: Ippudo East Village
whoAttended: John Smith, George Michael...
whenEatingOccured: 2017/10/07
whatWasEaten: Ramen
whyEatingOccured: -

whereEatingOccured: Aria
whoAttended: Maria Smith, John Willia...
whenEatingOccured: 2017/10/07
whatWasEaten: Italian
whyEatingOccured: Farewell party

whereEatingOccured: The Greek Tribeca
whoAttended: Anna Johnson, Cathrine ...
whenEatingOccured: 2017/10/07
whatWasEaten: Greek
whyEatingOccured: -

whereEatingOccured: Hillstone Restaurant
whoAttended: Irene Smith, Nick Miller, A...
whenEatingOccured: 2017/10/07
whatWasEaten: American
whyEatingOccured: -

(d) List of recognized restaurant outings.

whereEatingOccured: Ippudo East Village
whoAttended: John Smith, George Michael...
whenEatingOccured: 2017/10/07
whatWasEaten: Ramen
whyEatingOccured: -

Initiate Discussion ▼

whoInitiatedTheConversation: JohnSmith
whoWasIncludedInTheConversation: George Michael, Maria ...
whenWasTheConversationInitiated: 2017/10/06
whenIsTheProposedEvent: 2017/10/07
whereIsThePlanToGo: Ippudo

SendMessageOnMessenger
Related Items
Who is interested to go for dinner tomorrow?
Me, George, Maria

Make Reservation ▼

whoMadeTheReservation: John Smith
forWhoIsTheReservation: 4 people
whenWasTheReservationMade: 2017/10/07
whenWasTheReservationScheduled: -
whereWasTheReservationMade: OpenTable

Related Items
Your reservation confirmation for Ippudo East village
Opentable

WriteInCalendar ▲
Go to the Restaurant ▲

(e) Instantiation of a restaurant outing (1/2)

whereEatingOccured: Ippudo East Village
whoAttended: John Smith, George Michael...
whenEatingOccured: 2017/10/07
whatWasEaten: Ramen
whyEatingOccured: -

WriteInCalendar ▼
Related Items
Dinner with George and Maria

Go to the Restaurant ▼

whoWillGoTogether:-
whenToLeave: 2017/10/07
meansOfTransportation: Uber

Related Items
Your Saturday evening trip with Uber
Uber receipts

Attend EatingOut ▼
PostOnFacebook
Related Items
Saturday night out with my friends!
Tagged: George Michael, John Smith, Maria Smith

MakeAPayment ▲

whoPaid: John Smith
whenPaid: 2017/10/07
wherePaymentOccured: Ippudo East Village
whatWasPaymentAbout: Restaurant

(f) Instantiation of a restaurant outing (2/2)

Figure 2: *YourDigitalSelf* Screenshots.

declaratively defined (currently in an .xml notation). In addition, all the scripts/subscripts are parametric/generic in order to factor out the common parts of multiple scripts, so the same script can be used in different scenarios. For example, the makeAPaymentFor <restaurant> script is a parametric script of the form makeAPaymentFor <T> where T represents the type of thing the payment was for. This way, the same generic script can be used for the makeAPaymentFor <hotel> which can be part of a goingOnAVacation script. In addition, there is a declarative description of the evidence to search for the examining script, in the form of strong, mild and weak evidence. Based on the evidence, there is a declarative description of the clues to search in all the digital traces. Finally, there is a mapping file for all the *w5h* properties of the digital traces to the *w5h* properties of the script (e.g the whenPaid of a bank transaction is mapped to the whenEatingOccured of the EatingOut script).

To validate our approach, we presented a preliminary case study in [4]. The case study focused on identifying "going out to restaurant" events in the real-user data of three users. Our study showed the promise of our approach and confirms the need for integrating data from multiple sources to improve the accuracy of retrieval for each user. It also revealed that the users had very different variants of the "going out to eat" script, thus forcing us to focus in personalization in our current work.

The main implementation of our data integration and script instantiation [4, 5] takes place in the Application Manager layer (Figure 1), which communicates with the *w5h* -based data model layer.

3 DEMONSTRATION OVERVIEW

The demonstration will consist of two parts. In the first one, the audience will be guided through case studies involving a sample data set. In the second part, conference attendees will be invited to install the application on their own device to test the script instantiation algorithm on their own data.

3.1 Case Studies

Figure 2 shows screenshots of the *YourDigitalSelf* app that will be demonstrated. Figure 2(a),(b) show the settings screen, where users select the type of data they want to include and set up access to the various services. The list of available services is constantly updated as we develop the app; we will also provide users with the functionality to add specific service API access through open-source access to the app source code. Figure 2(c) shows a list of possible scripts that are being evaluated by the tool, for instance users can access a list of restaurants they have visited, trips they have taken, medical appointments, etc. As with services, we are continuously updating this list as we add more scripts. In addition, we are currently investigating the development of a script ontology through learning and/or crowdsourcing techniques. Figure 2(d) shows an example of results computed for the restaurant script. Each restaurant visit result has a set of *w5h* properties instantiated from one or more several digital traces, the details of the instantiation is shown in Figure 2(e-f), which show all the personal digital traces connected to the event, and which property they instantiated. For example, the system identifies that "John Smith, George Michael, and Maria" (and "Me") attended the ***Ippudo*** restaurant outing. This event was identified through instantiations of several sub-scripts: *"Initiate Discussion," "Make Reservation," "Write in Calendar," "Go to the Restaurant,...' '*. Each sub-script is instantiated by a different trace, e.g., an email, a messenger exchange, a Facebook post, a financial transaction. (A (sub)script may be instantiated by more than one piece of evidence.)

The potential values of *w5h* properties of the (sub)script are found through these traces, and need to be combined and corroborated. For instance, in the example of Figure 2, the identity of the dinner attendees can be inferred from the various piece of information: *John* made the reservation (strong evidence) for *4 people*, a pre-scheduling discussion involved *John, Maria and George* (weak evidence), a calendar event was set for dinner with *George and Maria* (strong evidence), and a Facebook post tagged *John, Maria and George* at Ippudo (strong evidence). Individually, each of these may not be enough to instantiate the corresponding *w5h* property with certainty (e.g., a reservation can be made for someone else, or canceled), but combined, they corroborate (or disqualify) each other.

We will show the audience the instantiation of various activity scripts from a sample user dataset. Details of the instantiations will be provided, including source provenance of the digital traces as well as scoring details for our instantiation algorithm.

3.2 CIKM Attendees' Participation

We will provide conference attendees with the opportunity to install the *YourDigitalSelf* application on their own (Android-based) mobile device. Attendees will be able to see the results of our instantiation algorithms on a variety of real-life scripts (restaurant outings, going to theater,...).

It is important to stress that *YourDigitalSelf* keeps all data on the user's device and therefore guarantees full privacy. However, if the conference attendee desire to help improve the *YourDigitalSelf* app, we will offer the option of aiding the evaluation of our techniques through feedback reporting capabilities. If the conference attendee choose to participate in the feedback, the tool will then report back to us, with the user consent, anonymized and aggregated quality control feedback measures. This will allow users to use our system over a long period of time, on their own data while maintaining the control and privacy of their data. Any data reported to us will be first shown to the user who will have the option of opting-out of the study at any time. The feedback will be used to refine our scripts and activity recognition as well as to incorporate personalized activity patterns in the *YourDigitalSelf* application in the future. In addition, we will provide tools for users to add access to their own services and build their own scripts.

4 RELATED WORK

There has been a lot of work on Personal Digital Assistant systems. Popular current systems include *Amazon Alexa, Siri,* and *Google Now.* These systems are aimed at giving users reminders based on their personal data, often with some commercial goal, and are limited to using information that is in the vendors' proprietary systems. In addition, these systems focus on prospective tasks: remembering to carry out tasks either based on a time or event trigger [8]; while our current application scenarios are centered around retrospective tasks: organizing past memories. Our data integration tool could however be used in conjunction with prospective approaches to personalize activities recognition. For instance, if the user *Alice* has been identified to book a cat-sitter every time she goes on a trip, when the *YourDigitalSelf* app recognizes that she is initializing a trip script (e.g., booking plane tickets, searching for restaurants in a different city,...) it can send her a reminder to book a cat-sitter.

A commercial tool closer in spirit to our approach is the "Inbox" Gmail interface (http:\inbox.google.com), which offers a feature to group emails by some pre-defined themes (scripts), such as *trips, purchases, finance.* However, these are only instantiated through information available by parsing the user's Gmail account, a significant limitation compared to our scripts, which use personal digital traces from a wide variety of sources.

REFERENCES

[1] W. Brewer. *Memory for randomly sampled autobiographical events*, page 21âĂŞ90. Cambridge University Press, 1988.

[2] A. Doan, A. Halevy, and Z. Ives. *Principles of Data Integration.* Elsevier Science, 2012.

[3] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16, 2007.

[4] V. Kalokyri, A. Borgida, A. Marian, and D. Vianna. Integration and exploration of connected personal digital traces. In *Proc. of ExploreDB'17 Workshop*, page 3. ACM, 2017.

[5] V. Kalokyri, A. Borgida, A. Marian, and D. Vianna. Semantic modeling and inference with episodic organization for managing personal digital traces. In *Proceedings of the 16th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE'17)*, pages 273–280. Springer, 2017.

[6] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 802–803, New York, NY, USA, 2006. ACM.

[7] P. Li, X. Luna Dong, A. Maurino, and D. Srivastava. Linking temporal records. *Proceedings of the VLDB Endowment*, 4(11):956–967, Aug. 2011.

[8] D. Schacter. *The seven sins of memory: How the mind forgets and remembers.* Houghton Mifflin, 2001.

[9] R. Schank and R. Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Psychology Press, 2013.

[10] Stanford Entity Resolution Framework. http://infolab.stanford.edu/serf/.

[11] D. Vianna, A.-M. Yong, C. Xia, A. Marian, and T. D. Nguyen. A tool for personal data extraction. In *Proceedings of the 10th International Workshop on Information Integration on the Web (IIWeb 2014)*, 2014.