A Tool for Personal Data Extraction

Daniela Vianna
DCS, Rutgers University
dvianna@cs.rutgers.edu

Alicia-Michelle Yong
DCS, Rutgers University
aliciamic.yong@rutgers.edu

Chaolun Xia
DCS, Rutgers University
cx28@eden.rutgers.edu

Amélie Marian
DCS, Rutgers University
amelie@cs.rutgers.edu

Thu Nguyen
DCS, Rutgers University
tdnguyen@cs.rutgers.edu

ABSTRACT

Digital storage now acts as an archive of the memories of users worldwide, keeping record of data as well as the context in which the data was acquired. The massive amount of data available and the fact that it is fragmented across many services (e.g., Facebook) and devices (e.g., laptop) make it very difficult for users to find specific pieces of information that they remember having stored or accessed. Unifying this fragmented data into a single data set that includes contextual information would allow for much better indexing and searching of personal information. Thus, we have developed a personal data extraction tool as a first step toward this vision. In this paper, we present this extraction tool, along with some preliminary statistics about personal data gathered by the tool for several users. The goal of the data analysis is to give a glimpse of what the digital life of a person may look like, and how it is currently partitioned across many different services; moreover, it reinforces the fact that it is not possible for users to manually retrieve, store and access their extensive digital data without the support of a personalized information management tool.

1. INTRODUCTION

Personal data is now pervasive as digital devices are capturing every part of our lives. Data is constantly collected and saved by users, either actively in files, emails, social media interactions, multimedia objects, calendar items, contacts, etc., or passively by various applications such as GPS tracking of mobile devices, records of utility usage, records of financial transactions, or quantified self sensors. Everywhere users go, in everything they do, a digital trace is produced, acting as a digital memory of their past actions, interactions, and whereabouts. The main challenges for personal information extraction lie on: (1) large amount of data, (2) scattered across devices and cloud services, and (3) stored in potentially very different formats in different places.

The richness of contextual information attached to the digital data can be of great help to users searching for information they remember having stored and accessed in the past. Answers to questions like what, when, where, who, why and how can guide a user during the search process, adding clues to events and actions surrounding the target data. For example, answering questions such as: when an email was sent, with whom was I talking, to what song was I listening, brings a handful of insights to the search process. It is important to highlight that a person's memory is notoriously unreliable, and fully trusting their recollection of contextual information can result in the loss of relevant search results. Therefore, context should be used as a flexible query condition, adding some degree of approximation or guidance to the search process.

A data extraction tool that accesses a variety of available services retrieving and storing users' data is a significant step towards the development of an individualized context-aware personal information management tool. This paper describe the current status of our personal information extraction tool as part of a larger project on personal information search and discovery. Members of the project have been collecting their own personal data and a careful analysis of this data collection reinforces our belief that it is hard for a user to manually manage and extract knowledge from their own personal data without the support of a personalized information management tool.

2. CHALLENGES

Creating a unified personal information management tool is not a trivial task. The first important step is the identification, retrieval, storage and modeling of all the data pertaining to a user. In this section, we will discuss the more relevant challenges encountered in the process of retrieving and storing a user's digital life to create a personal database that is robust, reliable and secure.

Most of a user's personal data is fragmented across multiple sources. Even in the best scenario where a user has complete control of his own data stored only on personal devices, it is challenging to keep track of every single bit of data stored over time, and it is even harder to remember exactly in which device the data is stored. The fact that personal data may not even be controlled by the user, since it can be spread across multiple third-party services, adds an extra challenge to the process of identifying and retrieving data. Although some web services provide access to data through programmatic APIs, retrieving the data from the sources can be tricky. The access to the APIs varies for each service and they are constantly being updated. Many common services do not export such APIs and require access via web query forms or outdated screen scrapping methods to retrieve the data. The extraction tool that we are proposing identified and implemented access to a variety of data sources, retrieving the decentralized data and storing it in a single database.

The heterogeneity of data storage formats across different devices and services presents a second major challenge. One possibility for addressing this challenge is to pre-process the data before storing it. However, this task, besides being time consuming, is prone to mistakes that could lead to missing important data. Pre-processing the data also requires the extraction tool to include deep knowledge of each data format available; this is a difficult process, especially given the rapid rate of changes in the services sourcing the data. To avoid these problems, we store the data keeping their original format in a NoSQL database that is already optimized for semi-structured data. Our current prototype uses MongoDB, a document-store system with a BSON encoding.

As the data is being retrieved, two new challenges arise: storage and privacy. In the last couple of years, the impressive growth in storage space while keeping costs low guarantees that tools as the one we are proposing can be implemented while imposing very little additional cost to the user. In the current state of our project, the personal data retrieved is being stored in the user's own hard drive. Even though this approach has some limitation in the sense that the data is only available locally, by storing it in the user's hard drive we can guarantee some clear privacy and security benefits. In the future, we plan to expand our current model to make the data available to the user from different devices and locations, but to adopt a more flexible approach, as is the case with personal clouds, requires careful handling of private data and support for user permissions.

3. EXTRACTION TOOL

In the process of creating a personal information database, the current version of our extraction tool identified and implemented access to a variety of data sources. The underlying personal data retrieved is stored in a flexible format that will allow us to perform data integration, search, and knowledge discovery. In this paper, we focus in describing all steps involved in the personal information extraction aspect of our project.

Figure 1 illustrates the main components of our project. As mentioned in the previous section, our main goal is to build a personal context-aware search tool that integrates a user's fragmented data set into a unified whole, and uses both data content and contextual information to support accurate personal information searches. The dotted boxes in Figure 1 around the Search and Knowledge Discovery components indicate that they are still in development and will not be detailed in this work. For the personal search, we aim to build an intuitive search tool based on the natural memory retrieval process, which relies on contextual cues to find past information. As for the knowledge discovery, our intention is to use knowledge discovery techniques to augment the extracted personal databases with individualized knowledge via query-base enrichment and rules.

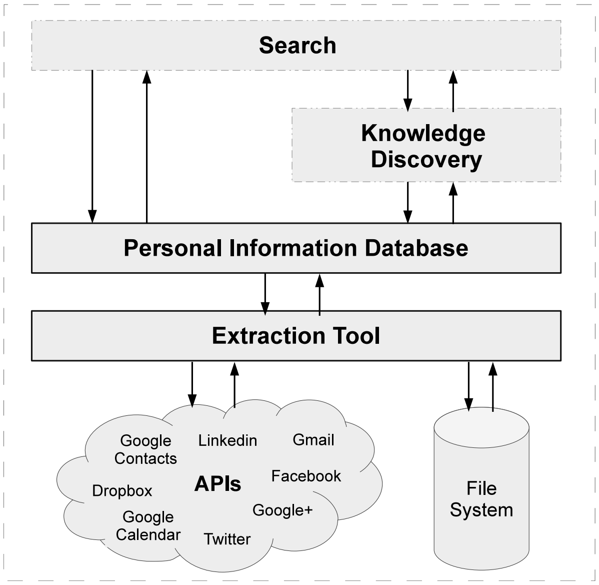


Figure 1: Architecture

The Personal Information Database (PID) illustrated in Figure 1 is responsible for storing not only the data retrieved by the extraction tool but also the information inferred by the Knowledge Discovery component. Besides that, the PID will hold all the information about a user necessary to access the APIs implemented – information such as access tokens are essential to authorize the tool to access data stored by third party applications.

The Extraction Tool follows the PID in the architecture illustrated in Figure 1. In this first version of our tool we are retrieving data from a wide range of sources: social data (Facebook, Twitter, LinkedIn, Google+), geolocation data (Foursquare, GPS), personal files (local file system, DropBox), Email (Gmail), Calendar (Google Calendar), Contacts (Google Contacts), and are planning on retrieving more. The data is currently being accessed through individual system APIs (when available), IMAP (for email retrieval), and a file system monitoring tool that we developed. The data collected includes content, structure and explicit and implicit context. In addition, we have started inferring additional contextual data. In particular, our implementation extracts time and location (when available) from every object it has processed and augments it with historical weather information retrieved from online sources.

Using a clean user-friendly interface, a user can authenticate and authorize the extraction tool to access their personal data for the range of services currently being offered. When a user registers a service using the extraction tool, a request is sent to the service API that, after the user has authorized it, will reply with an access token that is unique to the user and allows our tool to freely access the data. The access token is stored in the PID with all relevant information pertaining to that user. From now on, every time a user call the extraction tool to retrieve his personal data, the tool will query the PID for the access token and then will access the service API retrieving all the user new data since the last time a call was placed. It is important to highlight that a user's first attempt to retrieve data results in the tool trying to retrieve as much past data as possible as allowed by each API.

Another important part of the Extraction Tool illustrated in Figure 1 is the file system crawler, which was designed to run as a non-intrusive background process that keeps collecting every action a user performs in his own file system; information such as creation, edition and deletion of files are stored in the data collection together with all personal data pertaining to the same user. The crawler was designed to run indefinitely, unless the user wishes to stop the process.

With the exception of the file system crawler that was implemented using Java and JNotify – a file system events library for Java – all implementation was done using Python with the Django framework. Services are authenticated and authorized using Oauth2 and the data is accessed through APIs provided by each service. A preliminary version of our extraction tool was made available to the public in GitHub[1].

4. DATA ANALYSIS

In this section, we briefly discuss the services integrated in the current version of our extraction tool together with a description of the available data. Also, a statistical analysis of the data collected by members of the project over a period of one month will be discussed.

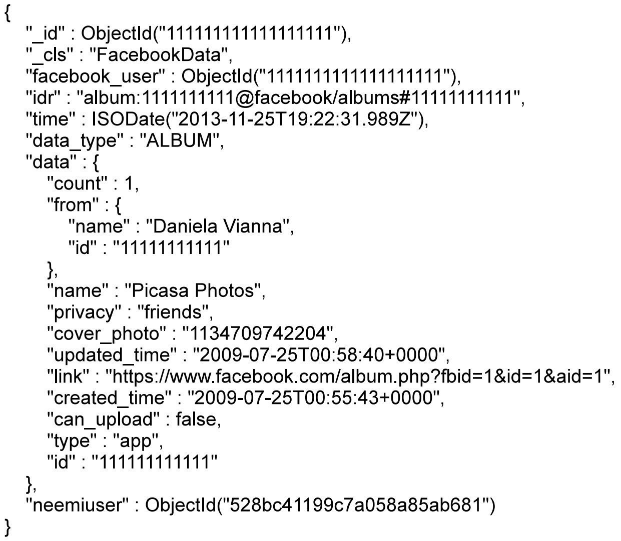
4.1 Personal Data

The variety of personal data available to be retrieved is enormous and new sources of data are constantly appearing; based on that, the extraction tool was built to easily integrate new services with their own data schema. As a starting point, our effort was channeled to selectively retrieve data from current popular services. Table 1 briefly describes the current state of our tool in terms of services and data retrieved.

Table 1: Services and data retrieved

Data Source	
Dropbox	files, folders
Facebook	feed, photo, album, checkin, event, friend, family, group, inbox, link, note, post, status, home, profile
Foursquare	badge, checkin, friend, photo, recent
Google Calendar	metadata, events
Google Contacts	contact, groups
Google+	people, activities, comments
Gmail	inbox, sent
Linkedin	connection, update, network, profile
Twitter	favorite, mention, friend, follower, timeline, retweet, msg received, msg sent, tweet
File system	created, modified, renamed, deleted
GPS	latitude, longitude, time

As mentioned in the previous section, the extraction tool retrieves and stores the data in BSON format using MongoDB. The data is not pre-processed in any way, i.e., the tool dumps the data preserving the original schema defined by the service from which it was retrieved. The absence of a unique pre-defined schema makes the tool robust to the very frequent changes in source APIs and export formats. Figure 2 illustrates a piece of data retrieved from the Facebook account of one of the authors. From this small piece of data we can extract information such as: time, user name, data type (Facebook album), album name and time the album was created and modified.



```
{
   "_id" : ObjectId("11111111111111111"),
   "_cls" : "FacebookData",
   "facebook_user" : ObjectId("111111111111111111"),
   "idr" : "album:1111111111@facebook/albums#11111111111",
   "time" : ISODate("2013-11-25T19:22:31.989Z"),
   "data_type" : "ALBUM",
   "data" : {
      "count" : 1,
      "from" : {
         "name" : "Daniela Vianna",
         "id" : "1111111111"
      },
      "name" : "Picasa Photos",
      "privacy" : "friends",
      "cover_photo" : "1134709742204",
      "updated_time" : "2009-07-25T00:58:40+0000",
      "link" : "https://www.facebook.com/album.php?fbid=1&id=1&aid=1",
      "created_time" : "2009-07-25T00:55:43+0000",
      "can_upload" : false,
      "type" : "app",
      "id" : "11111111111"
   },
   "neemiuser" : ObjectId("528bc41199c7a058a85ab681")
}
```

Figure 2: Data retrieved from the Facebook album of a user

Information such as time, location, text and people are frequently found in data from different sources. Besides those well know entities, the richness of the data and the possibilities that it offers in terms of how they are related and how they can be used to support a more robust search approach present a great stimulus to the study and development of a more personal context-aware search tool.

4.2 Statistics

In the current version of our extraction tool only text-based data and the metadata of multimedia objects are being retrieved. Table 2 shows monthly ranges over the amount of personal data retrieved by our extraction tool for members of the project. For a selected set of services, the table shows the number of objects retrieved – e.g., Facebook feed, photo, album –, the average size of those objects, and the total size of the data retrieved per service. Remember that the first call to the extraction tool results in retrieving as much past data as allowed by the requested service. Although storage space and cost do not impose a problem, the size of the data is impressive enough to make it impossible for a person to manually analyze the data.

Table 3 shows, for a subset of services and data types, the number of objects collected by a user during one month period. It is hard to estimate the amount of data produced by a single user, considering that the amount of personal data varies widely from user to user. The amount of objects in each collection can be an indication of how frequently a user accesses a service, showing the importance of a service

[1] `https://github.com/ameliemarian/DigitalSelf`

Table 2: Monthly Personal Data Statistics

Data Source	# Objects	Avg Size	Size
Dropbox	10-650	10Kb	0.05-16Mb
Facebook	500-2000	2Kb	0.75-3.8M
Twitter	700-1500	5Kb	3-10Mb
Gmail	1400-1700	200 Kb	310-390Mb
Google Calendar	20-100	2Kb	50-400Kb
Google Contacts	350-410	0.5Kb	230-275Kb
LinkedIn	120-140	1.5Kb	160-190Kb

to the user. The correlation between those different services represents a rich source of knowledge and highlights memory cues that could be essential to the success of a personalized search tool.

Table 3: Number of objects for different data types from Facebook and Twitter

Facebook		Twitter	
Data Type	# Objects	Data Type	# Objects
Feed	157	Timeline	662
Home	78	Tweet	16
Post	141	Mention	8
Album	1	Follower	10
Friend	145	Friend	20

5. RELATED WORK

A number of systems have been proposed for storing and retrieving personal data. Lifestreams [4] organizes desktop content in time-oriented streams. Stuff I've Seen [3] keeps an history of the web behavior of the user. Seetrieve [6] associate tasks, or user intentions, with user access patterns to aid in searches. Dataspaces [2, 1] connect personal data objects using semantic connections. Most of these systems assume that most data is available locally, or easily retrieved. In addition, while several do offer an integrated data model, their query models are typically keyword based, with sometimes one source of context used to aid the search. In contrast we envision a retrieval process that follows the memory process and uses all types of contextual cues.

Bell has pioneered the life-logging field with the MyLifeBits [5] project for which he has digitally captured all aspects of his life. While MyLifeBits started as an experiment, there is no denying that we are moving towards a world where all of our steps, actions, words and interactions will be recorded by personal devices or by public systems. SocialSafe [7] is a commercial tool that aims at extending Bell's vision for everyday users. The motivations behind SocialSafe are very close to ours, however SocialSafe currently only offers a keyword- or navigation-based access to the data (for a fee).

6. FUTURE WORK

In the future, we are planning to extend the personal data extraction tool to retrieve multimedia files instead of just their metadata. Also, the tool was implemented to make it easy to add new services without having to worry about specific details about different APIs, such as the structure of the data retrieved. We have been working on collecting browser history and information from applications of videoconference such as Skype and Google Hangouts.

The final goal of our project it is to build a context-aware search and a personal knowledge base. To achieve those goals we have been designing a context-aware data model that will be followed by the development of indexing and searching techniques, of a context-based query model, and finally, the implementation of a user search interface. We have also been exploring techniques for entity recognition and query-based enrichment for personal knowledge discovery.

7. CONCLUSIONS

In this paper, we presented an information extraction tool developed as part of a larger project on personal information search and discovery. The extraction tool is designed to collect heterogeneous data from a set of distributed data sources that varies from social networks to local filesystems. For privacy, the tool stores the retrieved data in the user's own hard drive by means of a personal information database. Preliminary analysis on the data collected by members of the project show the importance of personal management tools in dealing with large collections of personal data.

8. ACKNOWLEDGMENTS

This work has been partially supported by a Google Research Award. We thank Luan Nguyen for his contribution to the development of our filesystem crawler.

9. REFERENCES

[1] L. Blunschi, J.-P. Dittrich, O. R. Girard, S. Kirakos, K. Marcos, and A. V. Salles. A dataspace odyssey: The iMeMex personal dataspace management system. In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR'07)*, 2007.

[2] J.-P. Dittrich and M. A. V. Salles. iDM: A unified and versatile data model for personal dataspace management. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06)*, 2006.

[3] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff iÂŠ've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th International ACM SIGIR Conference (SIGIR'03)*, 2003.

[4] S. Fertig, E. Freeman, and D. Gelernter. Lifestreams: An alternative to the desktop metaphor. In *Conference Companion on Human Factors in Computing Systems*, CHI'96, 1996.

[5] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.

[6] K. Gyllstrom and C. A. N. Soules. Seeing is retrieving: building information context from what the user sees. In *Proceedings of the 2008 International Conference on Intelligent User Interfaces*, pages 189–198, 2008.

[7] Socialsafe. `http://socialsafe.net`.