Integration and Exploration of Connected Personal Digital Traces

Varvara Kalokyri
DCS, Rutgers University
New Brunswick, NJ 08904
v.kalokyri@cs.rutgers.edu

Alexander Borgida
DCS, Rutgers University
New Brunswick, NJ 08904
borgida@cs.rutgers.edu

Amélie Marian
DCS, Rutgers University
New Brunswick, NJ 08904
amelie@cs.rutgers.edu

Daniela Vianna
DCS, Rutgers University
New Brunswick, NJ 08904
dvianna@cs.rutgers.edu

ABSTRACT

A large number of personal digital traces is constantly generated or available online from a variety of sources, such as social media, calendars, purchase history, etc. These personal data traces are fragmented and highly heterogeneous, raising the need for an integrated view of the user's activities. Prior research in Personal Information Management focused mostly on creating a static model of the world (objects and their relationships). We argue that a dynamic world view is also helpful for making sense of collections of related personal documents, and propose a partial solution based on *scripts* – a theoretically well-founded idea in AI and Cognitive Science. Scripts are stereotypical hierarchical plans for everyday activities, involving interactions between mostly social agents. We augment these with hints of the digital traces that they can leave. By connecting Personal Digital Traces through scripts, we can build an episodic view of users' digital memories, which allow users to explore related events and actions in an integrated way. The paper uses the `Eating_Out` script for illustration, and ends with a report on the results of a case-study of applying a prototype implementation on real user data.

ACM Reference format:
Varvara Kalokyri, Alexander Borgida, Amélie Marian, and Daniela Vianna. 2017. Integration and Exploration of Connected Personal Digital Traces. In *Proceedings of ExploreDB'17, Chicago, IL, USA, May 14-19, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3077331.3077337

1 INTRODUCTION

Digital traces of our lives are now constantly produced by various connected devices, internet services and interactions. Our actions results 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 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" (PDT) are different from the traditional files that we are used to save and manipulate on a computer; they are typically (but not always) smaller, heterogenous, and accessible through a wide variety of different portals and interfaces, such as web forms, APIs or email notifications.

This paper takes steps towards organizing and summarizing such heterogeneous collections of PDTs by integrating connected traces into episodes, allowing users to explore their personal data, and allowing researchers to perform in-depth cross-services analysis and studies of user behaviors on various forms of services and media.

Traditionally, the field of Personal Information Management (PIM) has been concerned with gathering, integrating and querying of large collections of varied digital documents that are relevant to a specific person. The central focus of PIM systems such as Haystack [8], Semex [2] and OntoPIM [9] are the identification of relevant *objects* in the user's information space, and establishing their *inter-relationships*. Often, this is based on a domain or personal ontology. In contrast with this static view of information, we focus on a dynamic approach to the integration of PDTs, by providing a narrative to make connections between them. For example, a thread of messages concerning dinner, a confirmation of an OpenTable reservation at some restaurant, a Lyft receipt, a Foursquare check-in with photos, a credit-card payment, and a discussion on Facebook of a meal, makes much more sense as part of a narrative for going out to dinner, *if* they are appropriately related in time and location, and involve the same person(s).

The idea of a narrative to represent and explore personal data is supported by the notion of "episodic memory", originally introduced by Tulving (see [16], for example), which has been found to be a psychologically and physiologically distinct kind of memory for autobiographical events, encoding not just what happened, but times (*when*), places (*where*) and other contextual data. In fact, personal information can be unified and integrated through answers to the reporterial questions *who, what, when, where, why* and *how*

ExploreDB'17, Chicago, IL, USA
© 2017 ACM. 978-1-4503-4674-0/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077331.3077337

(the *w5h* questions). The work in this paper will use this *w5h* model, first described in [17].

We focus our approach on a predefined subset of common narratives, similar in spirit to the *scripts* introduced by Schank and Abelson [14]. A script is essentially a protoypical plan, "*a predetermined, stereotyped sequence of actions that defines a well-known situation*". An example of such a script would be "Caribbean travel for pleasure". The script would provide a description of possible "event flows" (e.g. make reservations, possibly get a visa or required immunizations, go to the airport, fly to the location, [...], eat at a local restaurant, [...], return, [...]). As with episodic memory, there are psychological studies supporting the notion of script [5]. The purpose of scripts is in our case two-fold: to *group and relate* some PDTs into a *script instance* that they relate to, and to extract information from these traces in a way that summarizes the specific episode, providing a higher level summary of personal information.

In this paper, we make the following contributions:

- We give a high-level description of our episodic scripts, allowing scripts to have their own properties and expressions describing valid sequences of actions, while adhering to the *w5h* organization.
- We describe a heuristic algorithm to find and combine evidence from PDTs for creating new script instances, filling their properties and those of their sub-scripts.
- We illustrate the above mechanism with a script for "going out to eat at a restaurant", including parts of a surrounding ontology of PDTs (which should include Facebook and Hangouts messages, Foursquare checkins, email, bank card transactions and Google calendar entries).
- In a case study, the above-mentioned algorithm has been implemented for the `Eating_Out` script, and we examine its performance for the purposes of recalling episodes of eating out at a restaurant.

2 RELATED WORK

PIM. The case for a unified data model for personal information has been made repeatedly. Haystack [8] offers basic linking, viewing and orienteering of a user's personal data. It focuses on individual information objects (mainly the metadata surrounding objects), and the relationships between them. Objects and relationships are modeled using RDF. Ontologies and some classes and properties are pre-defined; others, can be defined by the user.

Stuff I've Seen [3] indexes all of the information the user has seen (user's mail, web cache, personal files), and uses the corresponding metadata to improve search results. Most notably, Personal Dataspaces (e.g. [1, 2, 4]) propose semantic integration of data sources to provide meaningful semantic associations that can be used to navigate and query user data. Interactions in SEMEX [2] happens through a domain model of personal information. The ontology includes a set of classes and relationships. A collection of object-and-association extraction tools creates the data repository of objects and associations. In our model, we also rely on sub-classes and sub-associations, all defined based on the *w5h* properties.

The iDM [1] data model differs from Haystack and SEMEX, by representing heterogeneous personal information using graph structures that may be computed on demand.

The *OntoPim* project [9] explicitly suggests the use of a personal ontology as a way to organize personal data (and later connect this to performing office tasks).

As we shall see, in contrast to the above, our conceptual model for documents and objects focuses on the *w5h* properties and subproperties to help integrate the data less idiosyncratically, and uses property-chain inclusions to relate the properties of documents, actions, sub-scripts and scripts.

Processes and Plans. Since scripts are plans, and we want to recognize plan instances from the documents, the extensive literature on *plan recognition* is obviously relevant. A snapshot of this is available in [6]. The main difference is that these approaches tend to start from a description of a domain in terms of *planning operators*, while scripts are pre-compiled stereotypical plans.

Closely related to plan recognition is the areas of activity recognition, which often consider the problem of recognizing lower-level tasks, of which plans are composed, especially when these are signaled by sensors. The use of ontologies for this purpose is surveyed in [13], and a recent book [15], which combines the perspectives of the planning and activity recognition community, is an excellent survey of the field. The area of life-logging is quite similar, being based on sensors, and is surveyed in [7].

Our work is distinguished from most of the above efforts in the fact that there is massive concurrent execution of different kinds of scripts, that many digital traces are not part of any script, and that there are unobservable as well as exceptional variations performed in any script instance.

3 AN EPISODIC SCRIPT MODEL FOR PERSONAL DIGITAL TRACES

Personal Digital Traces can be grouped into related groups, or sequence, for the purpose of exploring or understanding past users actions. We focus our approach on episodic scripts.

We view scripts as stereotypical plans for everyday situations. They are prototypical in the sense that there are too many variations to capture them all ahead of time. Recall that our purpose for using scripts is to organize PDTs, abstract out relevant information, and help humans make sense of episodes. Therefore, though inspired by the work of Schank and Abelson [14], the details of our scripts will be different. Among others, they are not fully detailed plans, and can be more appropriately viewed as plan skeletons.

For our purposes, there are several relevant aspects to scripts: (i) summary information of the participants in the plan, and other descriptive properties, especially *w5h* aspects; (ii) the hierarchical decomposition into "invocations" of other scripts and primitive actions which describe the plan, together with restrictions on their ordering.

Due to lack of space, we will not present our full script language in this paper. We quickly illustrate our scripts ideas using the `Eating_Out` (aka "going out to eat") script. Such a script will have local properties with values that describe each instance of this script (e.g., `whoAttended`, `whenEatingOccurred`, `whereEatingOccurred`). Note that these properties are also organized, as much as possible, along the *w5h* dimensions.

The script **body** essentially describes the subgoals of the script plan, and constraints on the "flow of control" in achieving them.

In this case, after an action initiating the idea of going out on this occasion, there are discussions about when, who, what and where to eat, which can be carried out in parallel. Deciding when to eat in turn can be modeled by a script which shows exchanges of suggestions and discussions until agreement is reached.

Connecting scripts to PDTs. A connection must be made between the events in a script and PDTs, or more precisely the actions that give rise to them. For example, the reservation could be done via the OpenTable app or on its web page. If this information could be accessed via an API, it would provide almost certain evidence that `MakeRestaurantReservation` has been instantiated. Otherwise, an email from opentable.com which contains the word "reservation" in its Subject (indicating that a reservation has either been made or cancelled) provides strong evidence for `MakeRestaurantReservation` (although there is a possibility of it being advertising email, something we cannot determine without NLP, which we do not perform at the moment).

Our algorithm for connectiong scripts to PDTs starts by parsing the examining script's declarative definition with its *w5h* properties. Then, all the relevant PDTs are being retrieved by searching specific "clues" that provide evidence that the examining script has taken place. These clues are a list of "trigger words/phrases", whose occurrence indicates that a document is relevant with an instance of a particular script type. In order to find what clues to search for, one starts by identifying the *goal* sub-script/event(s). In the case of `Eating_Out`, it is the `attendEating` sub-script. One must then identify verbs that correspond to an occurrence of this event (e.g., "eat", "eat out" for `attendEating`). From this, a list of synonyms and hyponyms must be generated. In order to make this replicable for other scripts, we propose using *standard* sources of synonyms and hyponyms: WordNet, Cyc, ConceptNet5 [10–12].

The set of all documents D retrieved using these terms is preprocessed (i) by explicating/disambiguating information (e.g. terms like "tomorrow" or "Wednesday" are made absolute dates); (ii) by grouping certain kinds of documents (e.g. related email threads) into single individuals d in D. Each such individual leads to the creation of a candidate instance of the corresponding script. The *w5h* properties of each such document may fill some of the script instance (sub)properties. For example, a posting for a restaurant charge in a credit card provides evidence for the `attendEating` sub-script, together with information on its `whereEatingOccurred`, `whenEatingOccurred`, and one `whoAttended` filler (the card-holder).

One of the distinctive features of our system is the presence of multiple sources of evidence for the same script instance. In order to combine them, every top-level script needs to be analyzed to identify its *"key parts"*. The key parts are a subset of the *w5h* properties (or their sub-properties), which are saved in the local properties of the script, and which can help distinguish that particular instance of the script from others. For our `Eating_Out` example, the key parts are `whereEatingOccurred`, `whenEatingOccurred` and, to a lesser extent, `who`. The `why` and `what` local properties of this script are of secondary importance because they are rarely known without natural language processing, or because they would often lead to incorrect merging e.g. two instances of eating pizza need not be merged. When two instances of a script share an identical key part, they become candidates for *merging*.

	Alice	Bob	Charlie
Email/Messaging	56	52	21
Calendar	-	14	7
Financial Data	44	17	125 (49)
Location	9	-	-

Table 1: Number of objects/PDTs relevant to the `Eating_Out` script per source per user

4 CASE STUDY: EATING OUT

As proof of concept, we implemented our scripts for the `Eating_Out` scenario, where the goal is to find, among users personal data, the instances of them eating at various restaurants.

4.1 Gathering Personal Data

Performing experiments on Personal Data is not a trivial endeavor due to the sensitive nature of the data and the difficulty in getting personal data sets for research purposes.

For evaluating the utility of our approach, we gathered six months of Personal Data from three users: Alice, Bob and Charlie, by using our Extraction Tool described in [17].[1] We extracted their data from four types of sources: messaging (e.g. email, Facebook messenger, Hangouts), calendaring (e.g. Google Calendar), financial transactions (e.g. bank and credit card statements), and location services (e.g. Foursquare, Facebook checkins).

Table 1 shows the number of objects that were identified as relevant to the `Eating_Out` script in the six-month data sets of our three users. Relevance was computed using keyword-based scoring for the Email/Messaging and Calendar sources. Financial data and Location relevance were derived from the metadata categories stored with the original data items (for instance, a credit card payment at a restaurant will identify the expense as "Restaurants"). We verified and in some cases corrected this information by cross-referencing the address of the business using the Google Maps API. Note that the fact that an object is relevant does not mean that it indeed was part of an `Eating_Out` event. For instance, Alice may have discussed a restaurant in messages with friends but not gone there, or Charlie may have bought food at a business categorized both as a supermarket and a restaurant.

A first observation is that the three users have different patterns, which is expected because of the highly individual nature of user behavior. Alice, for instance, uses mostly messaging apps to make plans, rarely emails, and does not record dinner plans in her calendar. In contrast, she is the only one of our three users who uses location-based data [2]. Charlie shares a credit card account with her spouse, therefore some of the 125 relevant financial data objects are not from her credit card (only 49 are), nonetheless we kept all of them for our analysis as she often eats out with her spouse and the additional objects may contain relevant information as to her `Eating_Out` events. This illustrates the fact that Personal Data is often shared and that parts of a user information may reside in

[1] Users were volunteers. To preserve user privacy, we did not keep user data, only aggregated statistics.

[2] We could presumably extract more location-based data using GPS trackers on users phones.

Alice	Bob	Charlie
63	21	116 (40)

Table 2: Number of identified Eating_Out events per user

somebody else's data set. The ramifications of this with respect to PIM, especially search and script identifications are not yet fully understood and will be the topic of future work.

4.2 Identifying Eating_Out events: the Golden Set

To evaluate the quality of the memory retieval process using our scripts, we need to identify all the instances of Eating_Out for each user, aka a golden set. The identification of this golden set a posteriori is difficult because we cannot expect our users to accurately remember every single instance of Eating_Out. In the future, we are considering asking users to journal their lives over a long period of time. It is however unclear whether the mere act of journaling would have an impact on the type of data found in the user's Personal data - would the user record more information as a side effect of journaling? - and lead to a case of observation bias.

Without a perfect golden set, we cannot accurately evaluate Recall. However, we asked each user to carefully go over the six month of recorded PDT and asked them to identify all data that pertained to Eating_Out events. Table 2 shows the number of unique such events present in each user's data set, as identified by the users themselves.

5 EXPERIMENTAL EVALUATION

We now report on our experimental evaluation for the Eating_Out script over the data sets of our three users Alice, Bob, and Charlie.

5.1 Experimental Settings

5.1.1 Scoring. As explained in Section 4.1 and shown in Table 1, we identify relevant objects for each source separately. Objects that are identified as being relevant to the same script instance are then merged together.

5.1.2 Evaluation Metrics. We report on several metrics:

- **Percentage of events retrieved:** We report on the percentage of all user-identified Eating_Out events (see Section 4.2) that were retrieved by our scripts, as a proxy for Recall.
- **Overall Precision:** We report on the overall Precision, measured as the percentage of identified script instances that correspond to actual eating out events.
- **Precision@k:** We also report on the percentage of top-k (based in merged scores) script instances that correspond to actual eating out events.

5.2 Experimental Results

5.2.1 Impact of Different Sources. Table 3 shows the overall precision and percentage of identified events retrieved by our script for our three users. A first observation is that the results clearly reflect the different behavior of the three users. Alice and Bob use email/messaging to make restaurant plans in a majority of cases (59% and 86%, resp. Table 3), but do not always have a financial record of the transaction (67% and 52%, resp. - which could reflect cash payments. In contrast, Charlie makes very few plans by email/messaging (18% if only considering her data, but 6% if also considering events that only appear in her husband's financial data), nor does she enter them in her calendar (15% or 5% when including her husband's data), but most of her outings result in financial transactions (68%, 89% when including her husband's data). These results show that not only looking at several sources of information to identify script instances for a given user is critical to identify user script instances, as the percentage of events retrieved increases with the number of sources considered; but also that any approach to retrieve user memories of events must consider several sources to adapt to the wide variety of user behaviors.

	Alice	Bob	Charlie
Email/Messaging	0.59	0.86	0.06 (0.18)
Calendar	-	0.29	0.05 (0.15)
Financial Data	0.67	0.52	0.89 (0.68)
Location	0.14	-	-
Email/Messaging + Financial Data	0.98	1	0.95 (0.85)
Calendar + Financial Data	0.67	0.76	0.95 (0.83)
Location + Financial Data	0.68	0.52	0.89 (0.68)
Calendar + Email/Messaging	0.59	0.86	0.11 (0.33)
Email/Messaging +Location	0.7	0.86	0.06 (0.18)
All Sources	1	1	1

Table 3: Percentage of events retrieved per (set of) sources, per user

	Alice	Bob	Charlie
Email/Messaging	0.66	0.33	0.33
Calendar	-	0.43	0.86
Financial Data	0.95	0.65	0.82 (0.55)
Location	1	-	-
Email/Messaging + Financial Data	0.75	0.32	0.69 (0.4)
Calendar + Financial Data	0.95	0.52	0.83 (0.59)
Location + Financial Data	0.96	0.65	0.82 (0.55)
Calendar + Email/Messaging	0.66	0.32	0.46
Email/Messaging +Location	0.7	0.35	0.33
All Sources	0.75	0.32	0.67 (0.41)

Table 4: Overall Precision per (set of) sources, per user

Table 4 shows the overall precision of identified script instances. In this case, it is interesting to note that the quality of information given by different sources vary. Financial data tend to be of high quality (note that both Bob and Charlie had false positives due to the fact that they ordered takeout or bought groceries at a business doubling as a restaurant; we counted these as false positives. If these were to be considered reasonable Eating_Out events, then the precision for the financial data would be higher), whereas email/messaging data, which depends on keyword matching for relevance, tend to be of lower quality.

Taken together, the precision and percentage retrieved results confirm the need for (1) merging information from multiple sources of personal data to improve the identification of script instances for a given user in memory-based retrieval tasks, (2) considering a variety of Personal Information sources to account for the different individual behavior of users.

5.2.2 Quality of the Returned Answers. The results above show that our scripts achieve good precision. However, retrieval systems typically return results in a ranked order, and users are expecting the first few results to be the most relevant.

We now look at the quality of the returned answers by evaluating the Precision@k metric. Figures 1 and 2 show the results for all three users (for Charlie we show two plots, one with only her data, and one including her husband's financial data). Alice and Charlie's results show high precision over all values of k when including all sources of data. Alice's financial data is of very high quality, but as seen in Table 3, it only exists for 67% of her `Eating_Out` events. By combining Email/Messaging and Financial data information, she is able to identify her `Eating_Out` events with high accuracy for all values of k. Charlie's results show similar patterns, with even higher financial data accuracy, but lower Email/Messaging accuracy. Bob's data has some surprising patterns, his financial data is not as accurate as expected, in fact including it lowers the accuracy of his results post the top-13. This is due to the fact that the categorization provided by financial provider is inaccurate in several of his transactions. We are working on providing better correctness checking for third-party categorization data (in addition to the Google Maps API verification we have already implemented).

Figure 3 shows the Precision@k of various sources and source combinations, per user. We can see that overall, when information from multiple sources is combined (bottom-row), the precision, especially for low values of k, which are the first instances returned to the users, is higher than that of considering sources individually (top-row).

6 CONCLUSION AND FUTURE WORK

We have presented a script-based approach to organize heterogeneous collections of personal digital traces, as well as related user activities and scripts. The latter are stereotypical sequences of actions that can be used to connect PDTs "generated" by them into semantically coherent episodes, and to extract from PDTs relevant summary information.

We discussed a case study of applying it to some real-user data concerning going out to eat at a restaurant, and showed that our approach confirms the need for integrating data from multiple sources to improve the accuracy of retrieval for each user. Multiple sources also help reveal different patterns of social media and web use.

Using our script approach to group related personal digital traces into memory episodes would help users explore and navigate through their digital memories, being able to ask such questions as: "What are all my expenses related to my SIGMOD 2016 trip?" or "What are the last times I had dinner with my friend Paul?".

More generally, we believe this focus on episodic integration of PDTs could be valuable for researchers investigating the behavior of users across platforms and services. In addition, recognizing and separating out stereotypical events could improve the overall performance of systems that study and manage PDTs, by allowing them to focus scarce resources, such as intensive natural language processing, on the less frequent situations. For example, in studying a messaging service it could be useful to segregate first conversations that are part of episodes for standard forms of entertainment.

Our case study implementation exposed some of the challenges of working with PDTs. :

- Entity resolution: Identifying entities for the *w5h* dimensions is challenging. While the identification of *who* and *when* objects is relatively accurate, the other dimensions need improvement. For instance, we used Google Maps as an oracle for identifying *where* establishments, with partial success.
- Ambiguity: metadata information can be incomplete or ambiguous. In addition, real-life entities may have multiple uses: how to differentiate between a take-out vs eating in a purchase, or a purchase at a food store that has a restaurant section.
- Text analysis: An interesting NLP problem is how to deal with PDT text, which is typically short, highly informal, and may have scant context (or very individualized context).
- User involvement: how to incorporate user desires and feedback to personalize and disambiguate PDTs.

REFERENCES

[1] Jens-Peter Dittrich and Marcos Antonio Vaz Salles. 2006. iDM: A Unified and Versatile Data Model for Personal Dataspace Management. In *VLDB'06*. 367–378.

[2] Xin Dong and Alon Y. Halevy. 2005. A Platform for Personal Information Management and Integration. In *CIDR'05*. 119–130. http://www.cidrdb.org/cidr2005/papers/P10.pdf

[3] Susan Dumais, Edward Cutrell, Jonathan J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. In *Proceedings of the 26th International ACM SIGIR Conference (SIGIR'03)*.

[4] Michael J. Franklin, Alon Y. Halevy, and David Maier. 2008. A first tutorial on dataspaces. *PVLDB* 1, 2 (2008), 1516–1517. http://www.vldb.org/pvldb/1/1454217.pdf

[5] James A Galambos, John B Black, and Robert P Abelson. 2013. *Knowledge structures*. Psychology Press.

[6] Robert P Goldman, Christopher W Geib, Henry A Kautz, and Tamim Asfour. 2011. Plan Recognition (Dagstuhl Seminar 11141). *Dagstuhl Reports* 1, 4 (2011), 1–22.

[7] Cathal Gurrin, Alan F Smeaton, and Aiden R Doherty. 2014. Lifelogging: Personal big data. *Foundations and trends in information retrieval* 8, 1 (2014), 1–125.

[8] David R. Karger, Karun Bakshi, David Huynh, Dennis Quan, and Vineet Sinha. 2005. Haystack: A General-Purpose Information Management Tool for End Users Based on Semistructured Data. In *CIDR'05*. 13–26.

[9] Vivi Katifori, Antonella Poggi, Monica Scannapieco, Tiziana Catarci, and Yannis E. Ioannidis. 2005. OntoPIM: how to rely on a personal ontology for Personal Information Management. In *ISWC Workshop on The Semantic Desktop*. 258–262.

[10] Douglas B Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38, 11 (1995), 33–38.

[11] Hugo Liu and Push Singh. 2004. ConceptNet - a practical commonsense reasoning tool-kit. *BT Technology Journal* 22, 4 (2004), 211–226.

[12] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[13] Natalia Díaz Rodríguez, Manuel P Cuéllar, Johan Lilius, and Miguel Delgado Calvo-Flores. 2014. A survey on ontologies for human behavior recognition. *Comput. Surveys* 46, 4 (2014), 43.

[14] Roger C Schank and Robert P Abelson. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Earlbaum.

[15] Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P Goldman. 2014. *Plan, activity, and intent recognition: theory and practice*. Newnes.

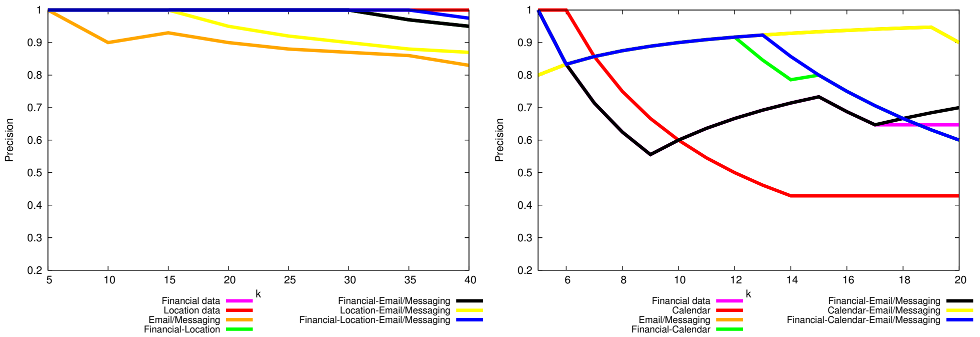

Figure 1: Precision@k for *Left:* Alice, *Right:* Bob

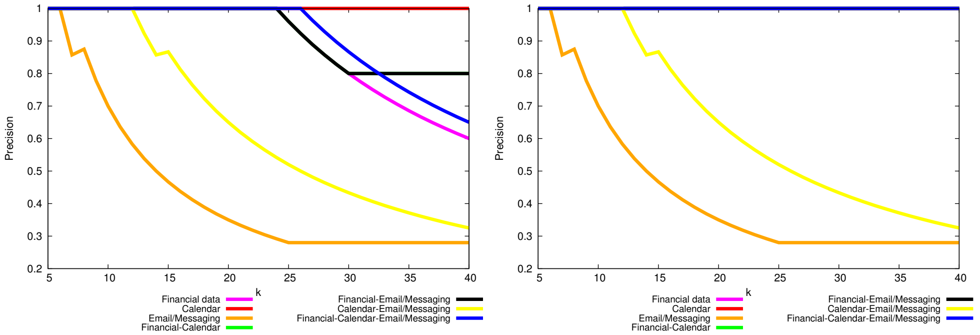

Figure 2: Precision@k for *Left:* Charlie, *Right:* Charlie, including her spouse's data

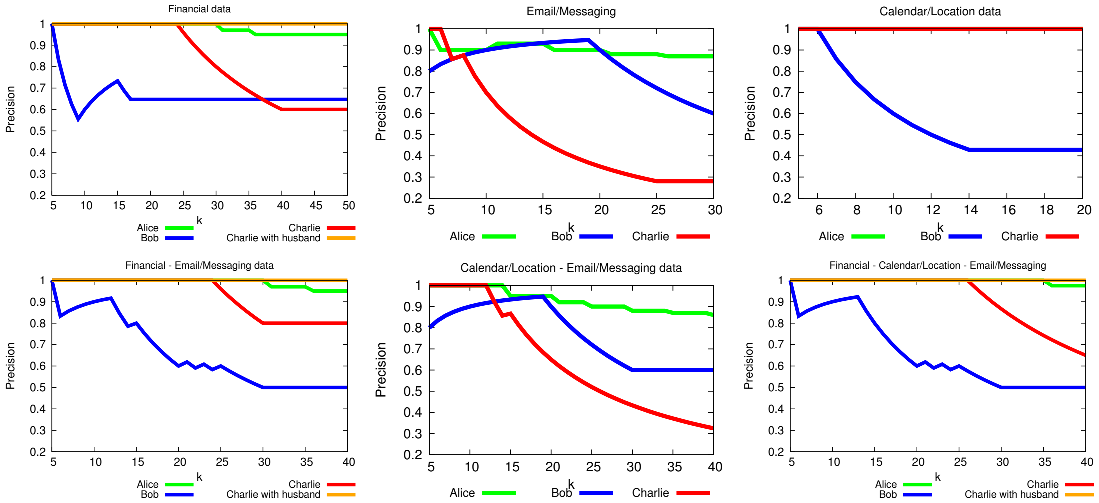

Figure 3: Precision@k for various source combinations: *Top-Left:* Financial Data, *Top-Middle:* Email/Messaging, *Top-Right:* Calendar/Location, *Bottom-Left:* Financial Data + Email/Messaging, *Bottom-Middle:* Email/Messaging + Calendar/Location, *Bottom-Right:* All Sources Combined

[16] E. Tulving. 2002. Episodic Memory: From Mind to Brain. *Annual review of psychology* 53 (2002), 1–25.

[17] Daniela Vianna, Alicia-Michelle Yong, Chaolun Xia, Amélie Marian, and Thu D. Nguyen. 2014. A Tool for Personal Data Extraction. In *Proceedings of the 10th International Workshop on Information Integration on the Web (IIWeb 2014).*