Searching Heterogeneous Personal Digital Traces

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

Varvara Kalokyri
DCS, Rutgers University
Piscataway, USA
v.kalokyri@cs.rutgers.edu

Alexander Borgida
DCS, Rutgers University
Piscataway, USA
borgida@cs.rutgers.edu

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

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

ABSTRACT

Digital traces of our lives are now constantly produced by various connected devices, internet services and interactions. Our actions result in a multitude of heterogeneous data objects, or traces, kept in various locations in the cloud or on local devices. Users have very few tools to organize, understand, and search the digital traces they produce. We propose a simple but flexible data model to aggregate, organize, and find personal information within a collection of a user's personal digital traces. Our model uses as basic dimensions the six questions: what, when, where, who, why, and how. These natural questions model universal aspects of a personal data collection and serve as unifying features of each personal data object, regardless of its source. We propose indexing and search techniques to aid users in searching for their past information in their unified personal digital data sets using our model. Experiments performed over real user data from a variety of data sources such as Facebook, Dropbox, and Gmail show that our approach significantly improves search accuracy when compared with traditional search tools.

KEYWORDS

Personal Digital Traces; Personal Search; Data Model

ASIS&T THESAURUS

Information Discovery; Knowledge and Information

INTRODUCTION

Digital traces of our lives are constantly being produced and saved by users, either actively in files, emails, social media interactions, multimedia objects, etc., or passively via various applications such as GPS tracking of mobile devices, web search records or quantified self-sensor usage. 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. These traces reflect a chronicle of the user's life, keeping record of where the user went, with whom the user interacted with, what the user did, and when. However, the large quantity of personal data available, and the fact that data is stored in multiple decentralized systems, in heterogeneous formats, makes it challenging for users to interact with their data and perform even simple searches.

Our goal is to give back to individual users easy and flexible access to their own data. Personal data is highly sensitive; consequently, privacy and ethical issues have to be considered while dealing with this type of information. The work discussed in this paper is developed as part of a series of tools to let user retrieve, store and organize their digital traces on their own devices (Kalokyri, Borgida, Marian, & Vianna, 2017a, 2017b; Vianna, Yong, Xia, Marian, & Nguyen, 2014), guaranteeing some clear privacy and security benefits.

Work in Cognitive Psychology (Brewer, 1988; Jones & Teevan, 2007; Schacter, 2001; Wagenaar, 1986) has shown that contextual cues are strong triggers for autobiographical memories. (Abowd et al., 1999) and (Dey, 2001) define context as any information that can be used to characterize the situation of an entity. This suggests that a natural way to remember and learn from past events is to include any pertinent contextual information when organizing and searching personal data. Personal information can be modeled, and indexed following six dimensions that mirror the basic interrogative words: what, who, when, where, why, and how. Each trace is a source of knowledge. For instance, a Facebook post may contain enough information to identify where a user went, what they did, who they interacted with, and when. Multiple traces, from the same or different sources, are often related to each other. The correlation between traces can be

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19–23 October, 2019

DOI: 10.1002/pra2.00022

identified through common information such as time and location. Even though multiple traces may share common information, they may have significantly different structures. This heterogeneity presents a major challenge. Thus, in this work, we are proposing a data model that can effectively represent this heterogeneous data, helping users find pieces of information again.

Search of personal data is usually focused on retrieving information that users know exists in their own data set, even though most of the time they do not know in which source or device they have seen the desired information. Current search tools such as Spotlight and Gmail search are not adequate to deal with this scenario as the user has to perform the same search multiple times on different services or/and devices rather than search over just a single service. Besides, traditional searches are often inefficient as they typically identify too many matching documents. In addition to the unified data model, we are proposing scoring and searching techniques that allow personal information search over data from multiple services and devices integrated in a unified data set.

In this paper, we make the following contributions: (1) a unified and intuitive multidimensional data model to link and represent heterogeneous personal digital traces; the model, called w5h, uses those six dimensions to unify features of each personal data object, regardless of its source (Section 3), (2) a frequency-based scoring methodology for searching personal digital traces; our scoring, named *w5h-f* is based on our multidimensional data model and leverages entities interactions within and across dimensions in the data sets (Section 4), (3) an implementation of our techniques, from data extraction, to entity recognition, classification and index structures, that will be used as the basis of our experimental evaluation (Section 5), and (4) a thorough qualitative evaluation of our proposed w5h scoring and search techniques, as well as comparison with two popular existing search tools, Solr (Apache Solr) and Spotlight (Apple, 2017), and techniques, TFIDF and BM25, on real data using both manually designed and synthetically generated search queries. Our results show that our scoring model results in improved search accuracy (Section 6). We discuss related work in Section 7 and conclude with future work directions in Section 8.

W5H DATA MODEL

We propose a data model that relies on the context in which personal data traces are created, produced and gathered to integrate heterogeneous traces into a unified data model that will support accurate searches. The proposed model, called w5h, was derived from the following observations: (1) personal digital traces are rich in contextual information, in the form of metadata, application data, or environment knowledge, and (2) personal digital traces can be represented following a combination of dimensions that naturally summarize various aspects of the data collection: *who, when, where, what, why and how*.

Our w5h model uses these six dimensions as the unifying features of each personal digital trace object, regardless of its source. Using these natural questions as the main facets of data representation will also allow the combination of our data representation with a natural and intuitive query model for searching information in digital traces. Listed below are some examples of dimensional data that can be extracted from a user's personal digital traces:

- **what:** messages, messages subjects, publications, description of events, description of users, list of interests of a user.
- **who:** user names, senders, recipients, event owners, lists of friends, authors.
- **where:** hometown, location, event venue, file/folder path, URL.
- **when:** birthday, file/message/event created-time, file/message/event modified-time, event start/end time.
- **why:** sequences of data/events that are causally connected.
- **how:** application, device, environment.

W5H SCORING MODEL

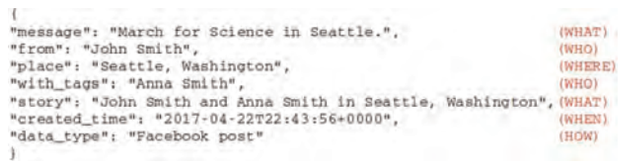

Figure 1. Simplified Facebook post classified according to the w5h model.

Figure 1 presents a digital trace from a Facebook post with each piece of information identified as belonging to one of the six dimensions proposed (what, who, where, when, why and how). Even though multiple digital traces come from different sources and have their own data schema, they can be unified using the six dimensions proposed in our w5h model. For instance, two separate traces that have John Smith or/and Anna Smith under the same dimension who (for example a Facebook image tagging Anna Smith, or a tweet mentioning John Smith), can be linked by our unified model. Details on the implementation of the dimension classification and entity resolution are given in Sections 5.2 and 5.3. The w5h model is used both to unify heterogeneous digital trace data from different sources, and to link digital traces using the six proposed dimensions.

The why dimension is not explored in this paper, but is the topic of related work (Kalokyri et al. 2017; Kalokyri, Borgida, & Marian, 2018). This dimension can be derived by inference and could be used to connect different fragments of data that derive from a common real-life task, or episode. For instance, the why value "March for Science Demonstration" may be inferred for the Facebook post in Figure 1 and could be used to connect to other related traces such as a message thread.

As pointed in (Wagenaar, 1986), users tend to remember their actions using the six natural questions; thus, using them to guide search is a logical approach. We now evaluate the potential benefits of the w5h model for integrating and searching personal data. Specifically, we propose a search mechanism that supports queries containing conditions along each of the six interrogative dimensions. Our proposed search relies on a novel frequency-based scoring methodology over the w5h data model, called *w5h-f*, that will be detailed in this section.

Scoring Methodology

To illustrate our query and scoring methodology let us consider the following search scenario: the user is interested in message(s) from John Smith or/and Anna Smith about the 2017 March for Science. We consider each digital trace to be a distinct object that can be returned as the result to a query.

DEFINITION 1 (Object in w5h Dataset). An object O in the data set is a structure that has fields corresponding to the 6 dimensions mentioned earlier. Each of these dimensions contains 0 or more items (corresponding to text, entities identified by entity resolution, times, locations, etc). The fields of an object O are accessed using functions O.get("who"), O.get("what"), etc.

Formal queries have the same structure as objects in the unified data set. In the example above, the query has three filled dimensions: March for Science (**what**); John Smith, Anna Smith (**who**); 2017 (**when**).

Given objects *Q* and *O*, *O* is considered as an answer to object *Q* treated as a query if it contains at least one of the dimensions specified in *Q*. In looking for (partially) matching objects to a given query, each dimension will be searched separately, and the results will be combined according to a scoring function, generating a rank-ordered list of candidates. The choice of scoring function can be application dependent. We propose our frequency-based scoring function, *w5h-f*, below.

w5h-f Scoring

Because personal digital traces are byproducts of users' actions and events, they are not independent objects. Our intuition is that the correlation between traces (objects) can be leveraged to improve the accuracy of search results. For example, if the March for Science query from Section 4.1 returns several potential matches, one from Alice Jones, and one from Bob White, we may want to score the one from Alice higher if she communicates more frequently as a group with the user, Anna Smith, and John Smith, than Bob White.

Our *w5h-f* scoring scheme uses the correlation between users (or entities) and how they interact over time to rank an object. Because we are focusing on personal digital traces, all the data articulates around a user. By analyzing the data collected by our Extraction Tool (Vianna et al., 2014) (Section 5.1), we observed a strong correlation between the user (owner of the data) and multiple users (*who* groups), through times (*who*, *when*), location (*who*, *where*) and data sources (*who*, *how*). For instance, in one of the datasets, 94.9% of the objects have more than 2 users (*who*), 95.7% of objects have at least one date (*when*), 99.9% of objects have content (*what*) and only 1.5% of the objects have location (*where*). Our scoring exploits those interactions and correlations by way of a frequency score.[1] Frequencies can be computed for individual users or group of users. They can be associated with multiple times, multiple data sources, and also with a set of locations. For example, from a set of emails exchanged between a group of users, we can extract the frequency (number of interactions) with which those users communicated, and in which time period those interactions occurred. In short, frequency expresses the strength of relationships, based on users, time, location and data sources (*who*, *when*, *where*, *how*).

To compute the frequencies across multiple dimensions, the frequency algorithm starts by retrieving a list of objects for each data source. For each object, the algorithm extracts groups of users, times and locations. Then, the following frequencies are computed:

- Frequency of each individual user: number of objects that mention a user in the who dimension.
- Frequency of a group of users: number of objects mentioning a group of users. If {a,b,c} is the group mentioned, frequencies of subgroups of {a,b,c}, e.g. {a,b} and {b,c}, are not counted.
- Frequency of each individual user at specific times: number of objects that mention a user at matching times. Time is normalized, so variations are also considered. For instance, a query searching for June, will match objects with time June 2016 and June 2017.
- Frequency of a group of users at specific times: number of objects mentioning the group at a specific time.
- Frequency of a location: number of objects that mention a location.

In addition to computing the frequencies per source, we also compute the total frequency of a user, group of users, times and locations by combining the individual results obtained for each data source. For simplicity, every time a user or group of users has an interaction, the frequency is increased by one; however, in practice, the algorithm allows us to weigh differently distinct types of interactions. For example, likes or comments on a Facebook post could be weighed differently, giving more relevance to interactions coming from comments than likes. Different roles, e.g. From and To in an email, can also be weighed differently.

[1] Our model is focused around personal digital traces and as such we included this specific group of correlations in our scoring. Other application scenarios could also benefit from our w5h, with other group and pairwise correlations highlighted in a dedicated frequency-based scoring. For instance, traces from weather sensors could have strong pairwise (where, when), or (where, how) correlations.

DEFINITION 2 (Similarity Score). Given a query Q, an object O, and the frequencies above, we define:

$$\begin{aligned} fscore(Q,O) = {} & f[g] + \sum_{u \in who} f[u] + \sum_{u \in who} fs[u] \\ & + \sum_{\substack{u \in who \\ dt \in when}} f[u][dt] + \sum_{\substack{u \in who \\ dt \in when}} fs[u][dt] \\ & + \sum_{\substack{u \in who \\ dt \in when}} f[g][dt] \\ & + \sum_{addr \in where} f[addr] \\ & + scorewhen(dt,O) + scorehow(s,O) \\ & + scorewhat(O) \end{aligned}$$

where g is the group of users in the who dimension of O, u is each user in g, dt is each time in the when dimension, s is a data source, addr is each location in the where dimension, f[*g*] *is the frequency of a group of users in the same object, f*[*u*] *is the total frequency of each user across all data services, fs*[*u*] *is the frequency of each user in the data source s of the object, scorewhen*(*dt*, *O*) = 1 *when the date dt from query Q matches object O; otherwise, scorewhen*(*dt*, *O*) = 0*, f*[*u*][*dt*] *is the total frequency of the user u in the time dt across all data sources, fs*[*u*][*dt*] *is the frequency of the user u in the time dt and data source s of the object, f*[*g*][*dt*] *is the total frequency of the group of user g in the time dt, f*[*addr*] *is the frequency of each location addr, and scorehow*(*s*, *O*) *is the score of an object O for a given source s: scorehow*(*s*, *O*) = 1 *when the service s from query Q matches object O; otherwise, scorehow*(*s*, *O*) = 0*. Lastly, scorewhat* (*O*) *is a text-based score for object O, using any chosen scoring function (e.g., TFIDF, BM25, …).*

The equation in Definition 2 assumes that a query *Q* has all 4 dimensions *who*, *when*, *where* and *how*; if a dimension does not exist in a query, the equation term corresponding to that dimension will be 0.

Let us consider the query Q_0 (*what*: March for Science; *who*: John Smith, Anna Smith; *when*: 2017), and the object *O1* illustrated in Figure 1 (Section 3). According to the *w5h-f* methodology, the object *O1* will have the following score:

$$\begin{aligned} fscore(Q,O1) = {} & f[g = John\,S., Anna\,S.] \\ & + f[u = John\,S.] + f[u = Anna\,S.] \\ & + fs[u = John\,S.] + fs[u = Anna\,S.] \\ & + f[u = John\,S.][dt = 2017] \\ & + f[u = Anna\,S.][dt = 2017] \\ & + fs[u = John\,S.][dt = 2017] \\ & + fs[u = Anna\,S.][dt = 2017] \\ & + scorewhen(2017,O) \\ & + f[g = John\,S., Anna\,S.][dt = 2017] \\ & + scorewhat\ \text{“}March\ for\ Science\text{”} \end{aligned}$$

Where s = Facebook.

SEARCH IMPLEMENTATION

We now discuss our search implementation in details.

Data Retrieval

To create a data set of personal digital traces, we use the extraction tool proposed in (Vianna et al., 2014) to identify and retrieve data from current popular services and sources of digital traces. The data retrieved is stored in its original format to avoid mistakes that could lead to missing relevant data. All the data collected by the tool is stored in MongoDB, a NoSQL database that is already optimized for semi-structured data, with the data from each service stored in its own collection. We are constantly adding and revising sources of personal digital traces; the current implementation includes emails services (Gmail), social networks interactions (Facebook, LinkedIn, Twitter), location services (GPS, Foursquare), file management (Dropbox, Local Filesystem), browsing data (Firefox, Chrome), financial data (Mint, bank accounts), calendars (Google Calendar).

In the next section we will present how the raw data retrieved can be parsed and mapped into the w5h model proposed in Section 3.

Classification

Having defined the w5h model (Section 3), it is still necessary to find an effective mechanism to translate the heterogeneous set of personal data into the six dimensions. The dynamic nature of data sources, especially the rapid rate of change in the service APIs, and the fact that new sources can be added into the extraction tool, also pose a challenge.

Digital traces have their own structures but most are retrieved in a semi-structured data format (typically JSON through APIs), or are extracted along with some metadata. We implemented parsers to represent the raw data from each source in the w5h model, thus unifying the data downloaded into a single data collection. The identification of data according to the six dimensions is done by analyzing the data available to be retrieved for each data source implemented and then building a dictionary of words/labels for each w5h dimension. Much of the classification is intuitive, for instance, the words From and To should be classified under the who dimension, while words Subject and Body should be classified as what.

We designed a machine learning multi-class classifier that automatically maps the raw data from each source into the w5h dimensions. The input data to the w5h classifier is a set of sentences and w5h labels. For instance, in Figure 1 (Section 3) each line corresponds to a sentence/label pair. Each sentence is then transformed in embedding vectors by a Word2vec algorithm, and labels are reshaped into one-hot encoded binary matrices. Architectures were built combining LSTM (Long Short-Term Memory) and Dense layers. Dropout was used in some architectures to reduce the complexity of the model with the goal to prevent overfitting. Parameters were evaluated using a 5-fold cross validation process to

estimate the performance of models. We use categorical cross-entropy as the training criterion (loss function); Adam optimization algorithm as the optimization algorithm for our models.

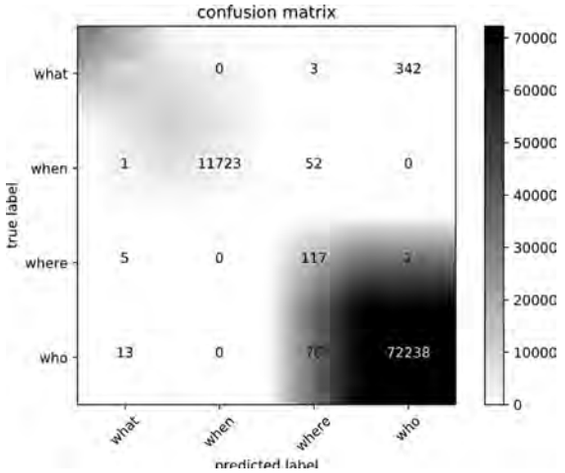

Figure 2. Confusion matrix with predictions for dataset *User 1*. The model was trained using dataset *User 2*.

The evaluation was conducted using the dataset *User 2* described in Table 1. The confusion matrix in Figure 2 shows the accuracy of the model for dataset *User 1* (Table 1), using the training data from *User 2*, with the true labels represented in the y-axis and predicted labels in the x-axis. All correct predictions are located in the diagonal of the table. The results indicate that a machine learning classifier can accurately translate dynamic and heterogeneous set of personal data into the w5h model. Our implementation uses the classifier to translate raw data into the w5h model and does not require user intervention.

Entity Resolution

Our scoring technique (Section 4) relies on frequency scoring of the same entity across objects. To make this possible, we need to identify separate instances of the same entity in data traces coming from the same sources, and across sources. For instance, the same person may appear in different services using variations of their names and email addresses. The impact of entity resolution on search performance will be discussed in Section 6.

Entity Resolution for the who Dimension

We use the Stanford Entity Resolution Framework (SERF), a generic open-source, infrastructure for Entity Resolution (ER) (Stanford), to identify entities. Using SERF person entities are identified and grouped in final entities that are stored in MongoDB in a separate collection.

Entity Resolution for the where Dimension

To disambiguate and match location data, we used Google Geocoding, Google Places API and SERF. We start by using Google Maps to disambiguate places that appear under different names and to augment the existing data, then we rank all addresses returned by a Google Maps search using a *tf* (term frequency) function computed based on the user's data set. When there are no results, we use location information from other related digital traces. We then use SERF for deduplication and record linkage for locations that have the same geocoded address information or geographical coordinates (longitude, latitude).

EXPERIMENTAL EVALUATION

We now evaluate the efficacy of the *w5h-f* search approach by comparing its performance with two popular existing search tools, Solr (Apache Solr) (using different scoring methodologies: TFIDF, BM25, and field-based BM25), and Spotlight (Apple, 2017). In this section, we first describe our evaluation methodology. Then, we explore the accuracy of the search approach for a set of search scenarios manually designed to be representative of possible user queries. Finally, we explore the accuracy of the search approach using a much larger set of synthetically generated searches.

	User 1		User 2	
Data source	**#Objs**	**Size**	**#Objs**	**Size**
Facebook	1493	9Mb	2384	19Mb
Gmail	1136	107Mb	10926	1Gb
Dropbox	-	-	573	32Mb
Foursquare	-	-	55	59Kb
Twitter	-	-	2062	10Mb
Google Calendar	2	9Kb	209	389Kb
Google+	1	1Kb	102	343Kb
Google Contacts	157	158Kb	427	430Kb
Total	2789	116Mb	16738	1.4Gb

Table 1. Personal data sets for two users.

Methodology

Data Set

There is a dearth of synthetic data sets and benchmarks to evaluate search over personal data. This challenge has only been exacerbated by the recent explosion in the amount of personal digital traces, as well as the varied services that create, collect, and store them. Thus, we perform our evaluation using a real data set collected by our extraction tool (Vianna et al., 2014) for two users.

Table 1 shows two real user data sets along with the number and size of objects retrieved from different sources over different periods of time. These two data sets will be used to evaluate the w5h scoring approach proposed in Section 4.

Search approach	Query description	Rank
Scenario 1 – search target: a Google+ post about SIGIR 2013 posted by Ashley in 2013		
Spotlight	MDContent:SIGIR, MDAuthors:Ashley, MDCreationDate:2013	2 - 14
TFIDF	SIGIR, Ashley, 2013	11
BM25	SIGIR, Ashley, 2013	12
Field-based BM25	who:Ashley, what:SIGIR, when:2013	8
w5h-f	who:Ashley, what:SIGIR, when:2013	5
Scenario 2: search target: a photo of a cat posted on Facebook by Katie in March 2012		
Spotlight	MDContent:photo, MDContent:cat, MDAuthors:Katie, MDCreationDate:2012-03	2-2964
TFIDF	Photo, cat, Katie, 2012-03	5468
BM25	Photo, cat, Katie, 2012-03	9106
Field-based BM25	what:photo, what:cat, who:Katie, when:2012-03	65
w5h-f	what:photo, what:cat, who:Katie, when:2012-03	13
Scenario 3: search target: a Facebook photo of Anna taken in Campos		
Spotlight	MDContent:Photo, MDContent:Anna, MDContent:Campos	2 - 3169
TFIDF	Photo, Anna, Campos	17
BM25	Photo, Anna, Campos	43
Field-based BM25	what:Photo, who:Anna, where:Campos	1
w5h-f	what:Photo, who:Anna, where:Campos	1

Table 2. Representative search scenarios targeting information stored in a user's personal data set.

Evaluation Techniques

Solr. Solr (Apache Solr) is a popular open source full-text search platform from the Apache Lucene project. For the experiments in this section, we integrate all data retrieved by the extraction tool, from each different data source, in a unified collection. This approach allows user to search for information across the entire set of retrieved digital traces, which is already a significant step forward from the current state. We consider three different scoring methods in conjunction with Solr: TFIDF, BM25, and field-based BM25 where the fields correspond to the parsing into the w5h model.

Spotlight. We also compare our search approach to Spotlight, the desktop search platform in Apple's OS X. Spotlight allows users to search for files based on metadata (Apple, 2017). This approach also works using the integrated raw (original) data. Each object in the evaluation data set is stored as an individual file in a machine running OS X Yosemite version 10.10.5. When possible, the following metadata is added to the files: MDAuthors (authors), MDCreationDate (creation date), MDChangeDate (content change date), MDCreator (content creator), MDFroms (path of a file). It is important to mention that Spotlight only ranks one item that it views as most relevant to a query. All other matching items are returned without ranking, typically organized by type of documents (e.g., email, pdf, etc.).

w5h-f. Our proposed approach relies on the six memory cues (what, who, when, where, why and how) to guide search. The *w5h-f* approach uses the data parsed according to the w5h model. The correlation between users/entities and how they interact over time through different services, including the frequency users communicate, is used to rank objects, as described in Section 4. *w5h-f* uses entity resolution, as described in Section 5.3, to disambiguate/link entities from different sources (e.g. Facebook, Gmail, Twitter...) in the data set.

Case Studies

We begin our evaluation by studying three manually created search scenarios designed to be representative of realistic user searches targeting different personal digital traces from the data set *User 2* described in Table 1. For each scenario, we compose one query for each of Spotlight, Solr (TFIDF), Solr (BM25), Solr (Fieldbased BM25) and *w5h-f* using the same

information. Query conditions are derived from information in the target objects, and all conditions are classified accurately along the dimensions within Spotlight, field-based Solr and *w5h-f*.

Table 2 describes the search scenarios, the corresponding queries, and the rank of the target object as returned by each search method. Note that the target objects are always found, since the queries are accurate, and all three search tools currently return all matching objects. When Spotlight does not return the target item as the 1st ranked result, we report the ranking as the range from 2 to the total number of returned items.

The results show that *w5h-f* achieves the best accuracy by always ranking the target object higher than or equal to Spotlight and Solr. The differences can be significant (e.g., scenarios 1 and 2), demonstrating that using memory cues to guide search can lead to improved search accuracy. We next discuss each of the search scenarios in more detail to show how differentiating between the dimensions, and using frequency information, helps to improve search accuracy.

In scenario 1, the user is searching for a data item containing information about the 2013 SIGIR Conference. The information was sent or posted by Ashley. In this scenario, identifying Ashley as who and 2013 as when allows *w5h-f* to rank the target object higher than all instances of Solr. When compared with Solr field-based BM25, using the same parsed data as *w5h-f*, the fact that *w5h-f* scoring function takes into consideration the frequency that Ashley communicated with the user during the year of 2013 using Google+, allows *w5h-f* to rank the target object higher than Solr. Spotlight was unable to leverage the same distinctions as *w5h-f* since the target object was not ranked number 1. Thus, Spotlight returned the target object as an unranked item among 13 other items.

Scenario 2 targets a photo of a cat sent or taken by Katie in March 2012. In this case, the classification of photo and cat as what and Katie as who allows *w5h-f* and Solr field-based BM25 to rank the target object much higher than Solr BM25, Solr TFIDF and Spotlight. Entity resolution in the who dimension and the scoring function based on frequency help *w5h-f* to rank the target object in the top 20.

Scenario 3 looks for a picture of Anna taken at a place called Campos. The good performance achieved by the w5h and Solr field-based BM25 approach is explained by the fact that those approaches were able to classify Anna under the dimension who and Campos under dimension where. Since Campos is a very common family name in the user database, the keyword search approaches ended up returning lots of documents matching Campos as location and also as a name.

Simulated Known-Item Queries

We now study a larger set of automatically generated known-item queries: search of personal data is usually focused on retrieving information that users know exists in their own data set. Considering the fact that personal data trace search is a known-item type of search, simulated queries can be automatically generated, using known-item query (Elsweiler & Ruthven, 2007) generation techniques such as the ones presented in (Kim & Croft, 2009), as detailed below.

For this set of experiments, we built two query sets, one using data set *User 1*, and one using data set *User 2* (Table 1). Both sets comprise 5 different groups of queries, each containing 1,500 queries for 250 different scenarios. Each scenario is automatically created by randomly choosing a target object from one of the evaluation data set. We then choose d dimensions, from which we randomly select v random values. We adapted the queries to each of our evaluation methods. The parameter d for each group is defined as following: Group 1, d = {what}; Group 2, d = {what, who}; Group 3, d = {what, who, when}; Group 4, d = {what, who, when, how}; Group 5, d = {what, who, when, how}. For Groups 1 to 4, v = 1. For Group 5, v = 2 to dimensions who and what, and v = 1 to dimensions when and how. We performed our experiments on both *User 1*, and *User 2* data sets and observed similar behaviors. For space reasons, we only report here on the results over the *User 2* data set.

Our evaluation resulted in the following observations on the impact of the multidimensional w5h data model, choice of text search function, entity resolution, and frequency scoring on the accuracy of the search results.

Methods	MRR	NDCG@10
Solr TF.IDF	0.2920	0.3384
Solr BM25	0.4742	0.5192
Solr Field-based BM25	0.4979	0.5428
w5h-f (no entity)	0.5632	0.5993
w5h-f	0.6119	0.6114

Table 3. MRR, NDCG@10 for Group 2 of queries.

Including pertinent contextual information when searching personal data can significantly improve accuracy. Tables 3 and 4 show the MRR (Mean Reciprocal Rank) and NDCG@10 (Normalized Discounted Cumulative Gain through position 10) of each approach, Solr TFIDF, Solr BM25, Solr field-based BM25, and *w5h-f*, for Group 1 - 5 of queries. If the target object has the same ranking as other matching objects, we report the median value of the range. Observe that all search implementations that use the data parsed according to the w5h model, Solr field-based BM25, and *w5h-f*, outperform the keyword-based approaches, Solr TFIDF and Solr BM25. These results show how valuable it is to use context (*w5h-f* and Solr field-based BM25) to find matching documents.

(a) Group 1

Methods	MRR	NDCG@10
Solr TF.IDF	0.1959	0.2304
Solr BM25	0.2127	0.2481
Solr Field-based BM25	0.2383	0.2712
w5h-f	0.2383	0.2712

(b) Group 3

Methods	MRR	NDCG@10
Solr TF.IDF	0.3580	0.4036
Solr BM25	0.5267	0.5619
Solr Field-based BM25	0.6117	0.6582
w5h-f	0.7072	0.7488

(c) Group 4

Methods	MRR	NDCG@10
Solr TF.IDF	0.3328	0.3925
Solr BM25	0.5357	0.5888
Solr Field-based BM25	0.6327	0.6765
w5h-f	0.7539	0.7931

(d) Group 5

Methods	MRR	NDCG@10
Solr TF.IDF	0.3772	0.4270
Solr BM25	0.5345	0.5924
Solr Field-based BM25	0.5769	0.6363
w5h-f	0.6514	0.7014

Table 4. MRR, NDCG@10 for groups 1, 3, 4, and 5 (Group 2 is in Table 3). Compared against w5h-f all the results are statistically significant (Wilcoxon signed-rank test).

The use of a more elaborated approach to search text data can positively impact the final results obtained by the w5h approaches. As previously mentioned, the what dimension in the w5h model is composed basically by content information comprising most of the text. *w5h-f* uses Solr field-based BM25 to score the what dimension. The impact of the text search using Solr field-based BM25 versus Solr TFIDF and Solr BM25, can be seen in Table 4 (a), which presents MRR and NDCG@10 for Group 1 of queries (queries have only the what dimension). We can observe that Solr field-based BM25 and *w5h-f* use a more efficient approach to search and score text data than Solr TFIDF and Solr BM25. Note that since Group 1 has only one textual dimension in the query, the *w5h-f* is equivalent to the underlying text-based scoring approach for the what dimension; field-based BM25 in our implementation. The results show that the adoption of a field-based text search for the what dimension leads to better results.

Being able to disambiguate/link people from different sources of data can significantly improve the accuracy of search. To analyze the importance of the entity resolution phase presented in Section 5.3, we created a group of queries (Group 2) composed by values from the who and what dimensions. The results, for the data set *User 2*, are illustrated in Table 3, with *w5h-f* approach being superior when using entity resolution, compared with an implementation of *w5h-f* that does not use entity resolution.

Including frequency information as part of the scoring results in significant improvements. Tables 3 and 4 show that *w5h-f*, which uses our proposed frequency scoring (Section 4), consistently outperforms Solr field-based BM25, which also relies on the w5h model (Section 3) but does not consider frequency. This shows that taking into consideration the correlation between dimensions while scoring an object improves the search accuracy.

Our evaluation shows that using a tailored frequency based multidimensional scoring approaches yields significant improvements in search accuracy over personal digital traces where the desired search outcome is a specific known object.

RELATED WORK

The case for a unified data model for personal information was made in (Karger, Bakshi, Huynh, Quan, & Sinha, 2005; Xu, Karlsson, Tang, & Karamanolis, 2003). deskWeb (Zerr, Demidova, & Chernov, 2010) looks at the social network graph to expand the searched data set to include information available in the social network. Stuff I've Seen (Dumais et al., 2003) indexes all of the information the user has seen, regardless of its location or provenance, and uses the corresponding metadata to improve search results. Most notably, Personal Dataspaces (Blunschi et al., 2007; Dittrich & Salles, 2006; Halevy, Franklin, & Maier, 2006) propose semantic integration of data sources to provide meaningful semantic associations that can be used to navigate and query user data. Our work is related to the wider field of Personal Information Management (Jones & Teevan, 2007), in particular, search behavior over personal digital traces is likely to mimic that of searching data over personal devices. Unlike traditional information seeking, which focuses on discovering new information, the goal of search in Personal Information systems is to find information that has been created, received, or seen by the user.

Bell has pioneered the field of life-logging with the project MyLifeBits (Gemmell, Bell, & Lueder, 2006) 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, and will generate a myriad of digital traces. Digi.me (digime) is a commercial tool that aims at extending Bell's vision to everyday users. The motivations behind digi.me are very close to ours; however, digi.me

currently only offers a keyword- or navigation-based access to the data; search results can be filtered by service, data type or/and date.

Other file system related projects have tried to enhance the quality of search within the file system by leveraging the context in which information is accessed to find related information (Chen, Guo, Wu, & Xie, 2009; Gyllstrom, Soules, & Veitch, 2007) or by altering the model of the file system to a more object-oriented database system (Bowman, Dharap, Baruah, Camargo, & Potti, 1994). YouPivot (Hailpern et al., 2011) indexes all user activities based on time and uses the time-based context to guide searches. Social context (users' friends and communities) is leveraged in (Smith, Barash, Getoor, & Lauw, 2008) for information discovery; similarly (Derczynski, Yang, & Jensen, 2013) uses temporal and location context to aid discovery in social media data. Our work integrates all these sources of contextual information and provides a unified complete model of context-aware personal data.

Contextual information has been considered in various computer science applications. Context-aware applications dynamically adapt to changes in the environment in which they are running: location, time, user profile, history. Bolchini et al. provide a thorough survey of context-aware models in (Bolchini, Curino, Quintarelli, Schreiber, & Tanca, 2007). Truong and Dustdar survey context-aware Web-Service systems in (Linh Truong & Dustdar, 2009). Context-awareness has become increasingly popular with the wide adoption of mobile devices. While the types of context these systems consider overlap with ours, the overall approach is different from ours, for instance a contextually-aware Information Retrieval system will use the current context (e.g., user location and time of day) to adjust search results (Shen, Tan, & Zhai, 2005). In contrast, we consider context as information that can be queried and used to guide the search.

CONCLUSION AND FUTURE WORK

We proposed and implemented a multidimensional data model based on the six natural questions: *what, when, where, who, why* and *how* to represent and unify heterogeneous personal digital traces. Based on this proposed model we designed a frequency-based scoring strategy for search queries that takes into account interactions between entities across objects to assist in the ranking of query results. Experiments over personal data sets composed by data from a variety of data sources showed that our approach significantly improved search accuracy when compared with traditional search methods. In the future, we plan on investigating several extensions to our work on searching personal data traces: (1) including topic modeling approaches over the *what* dimension to be able to correlate objects based on their contents, (2) designing dedicated indexes and search algorithms for search efficiency, (3) adding query relaxation rules to allow for approximate query matching, and (4) designing an aggregate query model where groups of objects (traces) can be returned together as a query answer (e.g., all the social media messages and pictures relating to a party), integrating the why dimension (Kalokyri et al. 2017) into our scoring framework.

REFERENCES

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999). Towards a better understanding of context and context-awareness. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing* (pp. 304–307). Springer.

APACHE SOLR. Retrieved from http://lucene.apache.org/solr/

Apple. (2017). *Spotlight*. Retrieved from https://developer.apple.com/library/content/documentation/Carbon/Conceptual/MetadataIntro/MetadataIntro.html

Blunschi, L., Dittrich, J.-P., Girard, O. R., Kirakos, S., Marcos, K., & Salles, A. V. (2007). A dataspace odyssey: The imemex personal dataspace management system. In *CIDR*.

Bolchini, C., Curino, C. A., Quintarelli, E., Schreiber, F. A., & Tanca, L. (2007). A data-oriented survey of context models. *ACM SIGMOD Record*, *36*(4), 19–26.

Bowman, C. M., Dharap, C., Baruah, M., Camargo, B., & Potti, S. (1994). A file system for information management. In *Proceedings of the International Conference on Intelligent Information Management Systems*.

Brewer, W. (1988). *Memory for randomly sampled autobiographical events* (pp. 21–90). Cambridge University Press.

Chen, J., Guo, H., Wu, W., & Xie, C. (2009). Search your memory! - an associative memory-based desktop search system. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 1099–1102). New York, NY.

Derczynski, L. R. A., Yang, B., & Jensen, C. S. (2013). Towards context-aware search and analysis on social media data. In *Proceedings of the 16th International Conference on Extending Database Technology, EDBT '13* (pp. 137–142). New York, NY.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing*, *5*(1), 4–7.

digi.me. Retrieved from https://www.digi.me.

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

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. (2003). Stuff I've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 72–79).

Elsweiler, D., & Ruthven, I. (2007). Towards task-based personal information management evaluations. In *Proceedings of the 30th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 23–30).

Gemmell, J., Bell, G., & Lueder, R. (2006). MyLifeBits: A personal database for everything. *Communications of the ACM*, *49*(1), 88–95.

Gyllstrom, K., Soules, C., & Veitch, A. (2007). Confluence: enhancing contextual desktop search. In *Proceedings of 30th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 717–718). ACM Press.

Hailpern, J., Jitkoff, N., Warr, A., Karahalios, K., Sesek, R., & Shkrob, N. (2011). YouPivot: improving recall with contextual search. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 1521–1530). New York, NY.

Halevy, A., Franklin, M., & Maier, D. (2006). Principles of dataspace systems. In *Proceedings of 25th ACM Symposium on Principles of Database Systems* (pp. 1–9). New York, NY.

Jones, W., & Teevan, J. (2007). *Personal information management*. University of Washington Press.

Kalokyri, V., Borgida, A., & Marian, A. (2018). Yourdigitalself: A personal digital trace integration tool. In *Proceedings of 27th ACM International Conference on Information and Knowledge Management, CIKM '18* (pp. 1963–1966).

Kalokyri, V., Borgida, A., Marian, A., & Vianna, D. (2017a). Integration and exploration of connected personal digital traces. In *Proceedings of the ExploreDB'17*, May 19, 2017, (pp. 3:1–3:6). Chicago, IL.

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

Karger, D. R., Bakshi, K., Huynh, D., Quan, D., & Sinha, V. (2005). Haystack: A general-purpose information management tool for end users based on semi structured data. In *CIDR* (pp. 13–26).

Kim, J., & Croft, W. B. (2009). Retrieval experiments using pseudo-desktop collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1297–1306).

Linh Truong, H., & Dustdar, S. (2009). *A survey on context-aware web service systems* (Vol. *5*, pp. 5–31).

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

Shen, X., Tan, B., & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th ACM SIGIR Conference Research and Development in Information Retrieval* (pp. 43–50). New York, NY.

Smith, M., Barash, V., Getoor, L., & Lauw, H. W. (2008). Leveraging social context for searching social media. In *Proceedings of the 2008 ACM Workshop on Search in Social Media* (pp. 91–94).

Stanford, I. *Stanford entity resolution framework*. Retrieved from http://infolab.stanford.edu/serf/

Vianna, D., Yong, A.-M., Xia, C., Marian, A., & Nguyen, T. (2014). A tool for personal data extraction. In *Proceedings of the 10th International Workshop on Information Integration on the Web (IIWeb)* (pp. 80–83).

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, *18*(2), 225–252.

Xu, Z., Karlsson, M., Tang, C., & Karamanolis, C. (2003). Towards a semantic-aware file store. In *Proceedings of the Workshop on Hot Topics in OS (HotOS'03)*.

Zerr, S., Demidova, E., & Chernov, S. (2010). deskweb2.0: Combining desktop and social search. In *Proceedings of Desktop Search Workshop, in Conjunction with ACM SIGIR 2010*, 23 July 2010. Geneva, Switzerland.