Carpé Data: Supporting Serendipitous Data Integration in Personal Information Management

1st Author Name
Affiliation
Address
e-mail address
Optional phone number

2nd Author Name
Affiliation
Address
e-mail address
Optional phone number

3rd Author Name
Affiliation
Address
e-mail address
Optional phone number

ABSTRACT

Author Keywords

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

In recent years, an unprecedented quantity and variety of information has been made available as structured data on the Web, through APIs, datasets, and data feeds. This information includes access to data previously hidden behind services and applications, such as retailer product catalogues, information not previously directly released to the public, such as open government data or records of personal financial transactions, and new kinds of data generated by emerging kinds of data sources, such as wearable and smartphone-based biosensors.

While a primary goal of opening up such direct access to information is the hope that end-user citizens will make more informed decisions pertaining to their health, wealth, or wellbeing [], use of this data beyond specialised app developers, journalists and other data specialists remains minimal. Our hypothesis is that the main impediment towards end-user access to this data is a mismatch between the capabilities of tools currently used to satisfy daily information needs, and those required to effectively browse, use and consume heterogeneous structured data from diverse sources. For example, while much of this data represents a many and varied set of things, activities, and descriptions at various granularities, most commonly used structured personal information management (PIM) tools either can only manage a small, fixed set of data types (such as digital calendaring tools and to-do list managers) or provide little or no support for structured data at all, such as text editors and sketching/drawing tools [].

If PIM tools were extended to enable end-users to make effective use of the emerging ecosystem of personal data, how would this impact personal information practice? This paper presents our initial results towards answering this question through a multi-method investigation of the personal data integration problem. We start with a small interview study examining information tasks people perform that rely on multiple sources of information, and how such needs are currently satisfied. Our results suggest that people prefer to rely upon multiple, diverse sources to singular integrated ones for a number of reasons, including coverage, reliability, and seeking out more suitable alternatives. Second, we conduct a analysis of various popular personal data sources available today to identify barriers to effective unification of data from each. We draw upon definitions and work from the database integration systems literature to characterise types of heterogeneity exhibited across data sources in three popular domains: social networking, shopping, and dining, finding that terminological heterogeneity dominates the simpler data feeds (including social networking and restaurant recommendations), while structural issues pervade the more complex schemas of online retailers' product catalogues.

Based upon the data integration needs revealed by people in the interviews, and the kinds of heterogeneity observed in the data feeds, we embarked on a user-centric design exercise towards an interface to facilitate simple, "light-touch" integration tasks of diverse, heterogeneous data sources. We describe our initial foray into this exercise, an interface we call DataPalette, which focuses on the most common types of heterogeneity to enable serendipitous "data mixing" - a form of simple integration sufficient to let people easily and effectively combine and compare information sources without the need to write code. Our usability evaluation of DataPalette reveals that most users were comfortable with and that this interface is effectively expedites the most sophisticated of the typical kinds of "data integration" tasks people need to perform while planning personal,

BACKGROUND

In this section, we touch on motivating work for DataPalette from three fields: database integration, peronal information management, and end-user mashups and toolkits.

The primary motivation for this work stems from field studies in personal information management, where letting users effectively consolidating and work with data from multiple sources has remained a challenge for over a decade. The problem of data fragmentation ([13], for example, arises, in part, from a failure of PIM tools to adequately support the effective consolidation of data across multiple applications, services and platforms (e.g., [1], [2]). In our own previous field study, (reference anonymised) we observed several different variations of manual "coping strategies" (or, if programmers, elaborate, custom-coded one-off solutions) that people devised to merge in data from external sources – be them online, offline, or other people. Similar findings from other studies include Voida et al.'s observations of volunteer coordinators' "homebrew databases", manually-mainted information assemblages concocted to handle the widely varied and heterogeneous data collection requirements such groups needed to coordinate their activities [15].

It became clear that effectively coping with data heterogeneity would be a central focus to letting people effectively draw upon multiple, arbitrary sources of data when consulting the theoretical data integration literature. Alon Halvey describes hetereogeneity as an inevitable by-product of the distributed processes that are used to author it:

> [The problem stems from the fact that] we are trying to integrate data systems that were developed for slightly (or vastly) different business needs. Hence, even if they model overlapping domains, they will model them in different ways. Differing structures are a byproduct of human nature ?people think differently from one another even when faced with the same modeling goal. [**?**]

In this paper, we take the position that there needs to be a difference in focus between the data integration approaches discussed in the databases literature and that of PIM, because of corresponding differences in requirements. In database systems, the overall objective is usually to allow multiple large databases to be accessed as if one. This requires essentially full mapping of all data types and objects and well-defined semantics concerning where and how data gets converted. The upfront costs of such complete integration are typically high, but are typically justified in the long-term, high-volume use cases imagined. In typical sensemaking scenarios of PIM, however, people might need to combine information from multiple sources for a one-off task; meaning that any upfront cost would be fully consume the time and effort a person has allocated for their task. Moreover, while a fully, uniform query capability might be nice, such capability is likely of less overall importance for a single quick task than simply being able to quickly and easily identify and cope with differences.

A number of prototype PIM systems have dmonstrated data integration capacities. For example, SEMEX [4] was a PIM environment that provided users a uniform query interface over the contents of their personal information environments, across heterogeneous files, folders and repositories. Similarly, Haystack was a personal information manager that allowed users to manipulate and managing their data in a uniform, consolidated manner [11]. Pertaining to each system's ability to integrate information from arbitrary new data sources, however, both systems generally required new sources to either already conform to a pre-specified set of ontologies, or have the user provide a manual specify such a mapping translation themselves.

Systems that have focused solely on the end-user-driven, interactive reconciliation of data heterogeneity have come out of the web "data mash-up" community. Yahoo Pipes! is a visual programming language aimed at letting novice programmers easily create data-transformation workflows, but remained rather difficult to use and poorly suited for one-off information tasks. Mash-up makers such as Intel Mashmaker [], QDEWiki, and TX2

DataPalette's approach is most directly inspired by "lightweight" visual, user-interactive data collection, consolidation and alignment tools such as Potluck [12] and Cards [7].

The Semantic Web community use ontologies to formally describe vocabularies for specific domains by defining axioms describing entities, instances of entities and the relationships that hold among them [3]. The axioms in an ontology can be either assertional or terminological, the former describes the schema of the entities and the latter the instances of those entities. Typically, these ontologies are marked up using OWL (Web Ontology Language) and Resource Description Framework Schema (RDF(S)). There has been an explosion of ontologies published online, many of which describe the same entities. These entities may or may not share the same entity names or properties. Therefore using data from heterogeneous ontologies can be difficult. The field of ontology matching can be split into two fields: schema and instance matching. Instance matching differs from schema level matching, because it aims to detect instances referring to the same real world object, while schema matching aims to find a set of mappings between concepts and properties in different ontologies.

For our purposes of allowing users to explore and compare data from different websites, we are particularly interested in aligning instances of the most common forms of data that people compare from website to website. Instance matching within this scope is in its infancy, some of its problems have been addressed by the record linkage problem in the database field [9, 16, 10]. However, there are new problems to address in this field, which are strongly related to schema matching, including structural heterogeneity and logical heterogeneity.

The approaches used to match instances typically use: natural language processing, using lexicons like Wordnet to identify synonyms, and word stems [8]; machine learning techniques [6]; probabilistic approaches [14] and comparisons of instances schemas structural and logical definitions [5]. In our case, this is problematic because information from the web does not always have a schema. In real-world scenarios, semi-automatic ontology matching approaches are favoured over automatic because reconciling the differences between ontologies that were designed for different purposes is not always accurate. Semi-automatic approaches use varying levels of support to aid a user, from prompting users to manually check flagged alignments, to suggesting possible alignments. Allowing users to compare instances across heterogeneous data is powerful because users can combine and evaluate multiple sources. In principle, automatic instance matching approaches would allow users to seamlessly browse this data, however for now and the foreseeable future such approaches cannot make use of personal knowledge and do not guarantee their alignments. Therefore in our work, we adopt a semi-automated matching approach, which will support the user to make alignments between web data and their personal data.

PRE-STUDY 1 - UNDERSTANDING HETEROGENEITY FOR EVERYDAY TASKS

Method

Our pre-study investigated the types of tasks people perform online that required heterogeneous data sources, and which tools they would use to organise a large social event involving 15 or more people. We interviewed 8 participants, using a general interview guide approach based loosely on a set of questions (see Appendix ?). This approach allowed both the participant and interviewer a degree of freedom in conversation and each interview could adapt to the participants experiences, while collecting the same general areas of information. Each interview was recorded and notes taken, the interviewer was trained so that they were knowledgeable about the importance of the study and to minimise bias.

Results and Discussion

The participant demographic consisted of 8 people, 7 of which were male and 1 female. They were split evenly into two age groups of 18-25 and 26-32. All but one participant used social network sites very regularly, and they either logged in multiple times a day, left a website page open or used dedicated application for their social networking sites updates. Those that used social networking websites primarily use Facebook. All of the participants used Twitter to listen to broadcasted tweets, and 6 out of 8 people regularly tweeted.

Question 1: Tasks

All participants had experience in using multiple websites to complete tasks. They listed example tasks such as shopping (including food, consumer electronics, hotels and flights), searching for jobs, choosing recipes, study and work. The biggest focus for shopping focused tasks was balancing price, quality and speed of delivery. In general, people did not not trust a single websites for reviews because they said that they could be biased and did not always review the features they were interested in. One participant said that video reviews of items were really important because they showed the aesthetics and scale of the item. In general for initial research into organising a task it was typical to Google keywords to find related websites to their tasks. They identified that they used different website for different jobs, such as manufacturers website for technical details, review aggregators for a range of opinions, and Google maps for location based decisions. Social networks were also identified to have different functions, such as different groups of people belong to each and share different types of opinions. One participant said that they shared their outlook and gmail calendars, but only for work events and did not record social activities on them.

Question 2: Large Social Event

All of the participants would organise a large social event by talking to their friends, about their ideas, preferences and recommendations. The tools they would use to acquire this information included face-to-face, phone calls, email, Skype, Facebook events. The method chosen depended on the time scale required to organise an event, in general if there was a short time scale people would speak in person or on the phone. Otherwise, people would set up a FaceBook event page to discuss ideas with their friends. If their friends were not on Facebook they would email them. All but one participant said they would not use Doddle to organise day and time of an event, because they felt that people did not fill in the form and it was more suit to organising work event not social.

Most people felt that being assertive and posting their decisions about the date and time on Facebook meant that organising events were more successful than trying to gather a consensus. They expected that their friends would voice any objections if they could not attend or like the restaurant or activity selected. They selected venues, locations, restaurants and activities based on their own knowledge, friends recommendations, Google maps, reviews through Googling and review sites such as tripadvisor. People stated that the cost, price and easy to get to locations were the most important factors when choosing a venue. They considered how to get there and looked up train or bus times on the Web. One participant organised carpools informally and at last minute via Facebook or text messages.

When organising an event the weather was not that important because it they did not trust its accuracy for events plan further in the future than one week. Also, their friends preferences in price and food choices were not a priority; they would try to be inclusive but only if it meant small changes to the plan. Choosing food and timings were often made on-the-fly, by either with the use of mobile applications or by walking past a location.

Prestudy 1 Summary of Findings

People said that they would like a website that could support the evolution of an event, that allowed them to post drafts of an event. They felt that organising an event should be a process that changes over time and did not need rushing. Three of the participants said that they would like a recommendation system for places to eat and activities. Another strong requirement was that it had to be ubiquitous so that everyone could use it because they wanted a single place to communicate with their friends. While Facebook was the most popular social networking site for organising events, half of the participants felt that it was not the best solution and would prefer a collaborative environment, which could be wiki based.

PRESTUDY 2 : TECHNICAL CHALLENGES OF INTEGRATING THE DATA

Method

Results

Analysis

Prestudy 2 Summary of Findings

WEBBOX INTERFACE - DESIGN OF THE DATAPALETTE

Design Goals, What are the problems observed needed to address What kind of interactions can be designed to solve them?

Interaction features multiple paths selection sameas interactions visualisation of partial results

Design process

METHODOLOGY

Participants

We recruited 20 participants (10 female) from the University of Southamptons international student mailing list. All of the participants spoke English, and for 12 of them, English was not their first language, however they had all passed the university English proficiency test. Predominately the participants were studying various levels of Computer Science, and 5 of which studied other subjects. All of the participants had previously encountered tasks requiring them to use and evaluate data from multiple websites.

Task Design

We designed a semi-structured lab-based study around four possible tasks shown in Table 1. The tasks were to select a restaurant for an event or university at which to study, out of six possible choices. We selected a shortlist of six because it is typical for users to narrow down their options to a limited selection before analysing the details of their final choices. The shortlist of universities and restaurants were taken from the top six ranked from the Complete University Guide (for task 1 and 2), Tripadvisor (for task 3) and Yelp (for task 4). The top six were chosen from these websites because they provide a realistic shortlist and are comparable in terms of rankings and properties. The web sites used as data sources were chosen because they are the most popular and contain comparable statistics for universities and restaurants.

Each task had an A and B condition. The A condition, the participant could only use WebBox to make their decision; and in the B condition the participant was given the choice to using any websites they desired and or spreadsheets in Microsoft Excel containing the data taken from the data sources to make their choice.

Each study consisted of two tasks, one A and one B condition. Each participant was given 10 minutes to complete a task. In order to select which tasks the participant was given, we used a latin square to permutate the tasks and conditions. In order to train the user on how to use our interface, they viewed a five-minute WebBox tutorial video introducing its basic functionality, how to use histograms and maps, and linking instances.

Data Collection

Each study was overseen by a facilitator and an observer: the role of the facilitator was to explain the studys protocol to the participant and to answer any questions; and the role of the observer was to observe without interacting with the user, and to take notes. During the lab study, we recorded the audio and the participants actions on screen. We asked the participants to speak out loud their thought process and what they were looking at, so that we could evaluate their reasoning behind their processes.

The studies lasted 40 minutes, on average. Both the facilitator and observer were trained on the purpose of the study, their role and how they could and couldnt influence the study. The studies were run over 4 days. At the end of each day the facilitators and observers met to discuss the studies data and to revisit protocols, if necessary.

Data Analysis

During the data collection we collated a spreadsheet summarising the participants gender, typical daily computer usage, the tasks allocated and basic actions performed during the task. After we had collected the studys data, we performed our data analysis.

TODO look further into types of data analysis....

Hypotheses

Interaction-method specific hypotheses

- h1.1 - Do people understanding the problem of coreference, and the need to reconcile coreference problems?
- h1.2. Does the ability to do drag + drop combination for co-reference reconciliation effectively solve this problem?
- h2.1- Do people understand the problem of structural differences in data?
- h2.2 - Does multipath selection let people effectively deal with this – and work with collections of heterogeneous items?

Are these interaction methods sufficient to perform common tasks involving heterogeneous data sources? Does supporting these simple techniques facilitate task completion?

RESULTS

Users who opted to use websites to make their decisions repeatedly encountered a number of difficulties and issues. In particular, there was significant confusion over the meaning of data points, for example the percentage-based rankings of university teaching by the Times Higher Education. There were numerous usability issues in using websites, particularly in searching for university courses and course requirements. Similarly, participants found it difficult to understand the locations of restaurants in cities that they were not familiar, and could not find a way to plot the shortlist of restaurants onto the same map using the websites. Overall, the users that spent the most time using websites exhibited significant frustration attempting to locate and compare different criteria (such as course entry requirements, or restaurant locations), and all ended up falling back to using the supplied spreadsheets.

One of the designs of the study was to produce a spreadsheet worksheet for each website, which meant that users had to switch between the worksheets to gather data from multiple sources. Different approaches to using the spreadsheets

Table 1. Tasks, choices and data sources

Task	Choices	Data Sources
1) Three universities they would apply to study Sport Science	Universities: 1) Loughborough, 2) Durham, 3) Exeter, 4) Edinburgh, 5) Birmingham, and 6) Bath	http://www.thecompleteuniversityguide.co.uk/, http://www.ucas.ac.uk/, http://unistats.direct.gov.uk/, http://www.timeshighereducation.co.uk/world-university-rankings/
2) Three universities they would apply to study History	Universities: 1) Cambridge, 2) London School of Economics and Political Science, 3) Durham, 4) Oxford, 5) St Andrews, and 6) University College London	http://www.thecompleteuniversityguide.co.uk/, http://www.ucas.ac.uk/, http://unistats.direct.gov.uk/, http://www.timeshighereducation.co.uk/world-university-rankings/
3) Restaurant they would book for 12 friends in Cambridge	Restaurants : 1) Alimentum, 2) Gardenia Restaurant, 3) The Cambridge Chop House, 4) Al Casbah, 5) Midsummer House Restaurant, and 6) Nando's Restaurant.	www.yelp.co.uk, www.tripadvisor.co.uk, http://ratings.food.gov.uk/, facebook and google plus.
4) Restaurant they would book for 12 friends in Glagow	Restaurants :1) Black Sheep Bistro, 2) Charcoals, 3) Number 16 Family Restaurant, 4) Fanny Trollopes Cafe, 5) Two Fat Ladies At The Buttery, and 6) Sapori D'Italia.	www.yelp.co.uk, www.tripadvisor.co.uk, http://ratings.food.gov.uk/, facebook and google plus.

were observed. Some users were keen to collate all data onto a single sheet, so that they could compare the values all at once. However, this approach was error prone, for example participants noticing too late that the ordering of rows was different on each sheet, so that their pasted values were incorrectly mapped. Others used pen and paper to collate the values, which was often faster and resulted in less observed frustration. While the majority of users did not attempt anything complex with Excel (in fact only one users used charts), there was some advanced use of Excel observed. One user explained that they did not like to look at numbers, and so coloured the cells of the types of restaurant, and the friends' preferences, and matched the colours to make their decision. Another user calculated their own weighted metric of which university to choose by averaging different metrics supplied by the data sources.

The most commonly observed use of WebBox was to collate the data on-screen, and compare multiple values directly. About half of the observed users actually used the instance combination feature of WebBox, with the other half able to successfully make decisions without using it. Those who tried to use it were largely successful, and were able to therefore use the global brush-highlighting to simultaneously compare different statistics. Most users instantly opened up all data sources into individual boxes, in order to determine which sources contained specific statistics. The result of this technique is that the users instantly filled their screen, and ran out of space for plots and maps.

A number of users admittedly introduced their own bias into the choices: "I prefer Indian food, so I will choose the Indian restaurant", and one participant spent time on Google Maps looking at the routes to each restaurant, and eliminating those that required bus transfers (the protocol did not mention routing/transfers at all).

DISCUSSION

In order to complete the tasks, the users followed a number of different processes. The most common process followed was *collation* of data, either on-screen or on paper. Some users were proficient enough to copy and paste data in the supplied spreadsheets from worksheet to worksheet, while others manually typed data value-by-value across worksheets. Others preferred to collate the data from screen to paper, in order to compare multiple data points. A similar process that some participants followed was to determine which values of a particular statistic were acceptable, and keep a paper tally of these next to the set of choices. They then made their choice based on the value of this tally.

Two contrasting approaches that participants followed was to use the data to rule out candidates, or to use the data to directly compare candidates at once. Our pool of participants exhibited both behaviours evenly, however in the condition when using the WebBox interface, significantly more comparison (and visualisation of values) was performed, with fewer participants using the "ruling out" process. A consequence of this was that when asked to justify their choices, users were able to verify their values on-screen immediately, whereas those who had ruled-out had to memorise their reasons, with mixed success – many participants had to look through the data sources to re-find the reasons why they has ruled out some choices. Therefore, the uses that were using WebBox were more confident in their choices, and knew exactly why they had chosen them, often pointing to the screens to show and verify each choice as they explained them.

CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

1. Bergman, O., Beyth-Marom, R., and Nachmias, R. The project fragmentation problem in personal information

management. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, ACM (New York, NY, USA, 2006), 271–274.

2. Boardman, R., and Sasse, M. A. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proc. CHI '04*, ACM Press (2004).

3. Borst, W. Construction of engineering ontologies for knowledge sharing and reuse.

4. Cai, Y., Dong, X. L., Halevy, A., Liu, J. M., and Madhavan, J. Personal information management with SEMEX. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM Press (New York, NY, USA, 2005), 921–923.

5. Castano, S., Ferrara, A., and Montanelli, S. Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics V* (2006), 25–63.

6. Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., and Halevy, A. Learning to match ontologies on the semantic web. *The VLDB Journal 12*, 4 (2003), 303–319.

7. Dontcheva, M., Drucker, S. M., Salesin, D., and Cohen, M. F. Relations, cards, and search templates: user-guided web data integration and layout. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, ACM (New York, NY, USA, 2007), 61–70.

8. Euzenat, J. An api for ontology alignment. *The Semantic Web–ISWC 2004* (2004), 698–712.

9. Fellegi, I., and Sunter, A. A theory for record linkage. *Journal of the American Statistical Association* (1969), 1183–1210.

10. Gu, L., Baxter, R., Vickers, D., and Rainsford, C. Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report 3* (2003), 83.

11. Huynh, D., Karger, D., and Quan, D. Haystack: A Platform for Creating, Organizing and Visualizing In-formation Using RDF, 2002.

12. Huynh, D., Miller, R., and Karger, D. Potluck: Data mash-up tool for casual users. *Web Semantics: Science, Services and Agents on the World Wide Web 6*, 4 (Nov. 2008), 274–282.

13. Jones, W., Karger, D., Bergman, O., Franklin, M., Pratt, A., and Bates, M. Towards a unification and integration of pim support, 2005.

14. Suchanek, F., Abiteboul, S., and Senellart, P. Paris: probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment 5*, 3 (2011), 157–168.

15. Voida, A., Harmon, E., and Al-Ani, B. Homebrew databases: complexities of everyday information management in nonprofit organizations. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, ACM (New York, NY, USA, 2011), 915–924.

16. Winkler, W. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*, Citeseer (1999).