An activity based data model for desktop querying (Extended Abstract)*

Sibel Adalı[1] and Maria Luisa Sapino[2]

[1] Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA,
sibel@cs.rpi.edu,
[2] Università di Torino, Corso Svizzera, 185, I-10149 Torino, Italy
mlsapino@di.unito.it

1 Introduction

With the introduction of a variety of desktop search systems by popular search engines as well as the Mac OS operating system, it is now possible to conduct keyword search across many types of documents. However, this type of search only helps the users locate a very specific piece of information that they are looking for. Furthermore, it is possible to locate this information only if the document contains some keywords and the user remembers the appropriate keywords. There are many cases where this may not be true especially for searches involving multimedia documents. However, a personal computer contains a rich set of associations that link files together. We argue that these associations can be used easily to answer more complex queries. For example, most files will have temporal and spatial information. Hence, files created at the same time or place may have relationships to each other. Similarly, files in the same directory or people addressed in the same email may be related to each other in some way. Furthermore, we can define a structure called "activities" that makes use of these associations to help user accomplish more complicated information needs. Intuitively, we argue that a person uses a personal computer to store information relevant to various activities she or he is involved in. Files may be related to activities either directly or indirectly with some degree of relationship. In this paper, we define a simple model of an activity and show the types of queries that can be answered using the activity model. Our model assumes that activities can involve files that are related to each other in many different ways: a period of time that may contain disjoint intervals, different locations, a group of people that we interact with and various combination of these types of associations. Furthermore, files may be related to multiple activities independent of their participation in one activity. Finally, our model aims to find the best indicators of an activity for a specific user and computer based on the data provided by that user.

* This work was supported by the National Science Foundation under grants EIA-0091505 and IIS-9876932.

2 Activity based querying

As a movitating example, suppose the user wants to find the photo of the Panda from her trip to the zoo and her photos do not have the necessary tags. It is possible to search for this information by first finding the time frame for the specific trip to the zoo by using a keyword query for all the relevant files and then limit the search to files created or photos taken at this time frame. Similarly, it is possible to limit searches to relevant people, directories based on the user's needs and find information by following associations known to her. In this case, we are able to find specific information and at the same time follow the links to browse the related information along different dimensions. This is similar to the way we recall information that we do not remember. To accomplish this, the system simply needs to show the relevant associations for any searched query.

To facilitate this type of querying, we define the notion of an activity as follows: Suppose $\mathcal{O}$ refers to the universe of objects that could be stored in the computer. Then, an *activity* actF is defined as a function $\mathsf{actF}_{\preceq} : \mathcal{O} \rightarrow D_\tau$ where $\tau = (D_\tau, \preceq)$ is any partial order. Intuitively, an activity is an outside event that triggers the use of a computer and the creation or use of data. Examples of professional activities that an academician may be involved in are publishing papers at conferences or journals, sending proposals, teaching classes, etc. Examples of personal activities may be taking trips, participating in sportive activities and personal gatherings, etc. We are not interested in modeling the meaning of these activities, but how they cause the creation of data objects for this specific user. For example, for a trip to visit friends or family, pictures taken at that trip, emails and web site visits corresponding to purchase of tickets and email correspondence with friends can all be considered relevant to the trip. These in fact model different aspects of the trip. For a conference, we might also create documents such as papers and presentations in addition to the files associated with a trip. To define an activity, we assume the user defines an *activity schema* actS as an ordered list $\mathsf{actS} = \langle\, lf_1 \ldots lf_k \,\rangle$ of logical formulae lf_i constructed from predicates defining the "where", "when", "what" type of constraints with possible crisp or fuzzy semantics. The activity actF defined by the above schema is then given by:

$$\mathsf{actF}(o) = \begin{cases} min\{i \mid o \models lf_i\} & \text{if } \exists i.(1 \leq i \leq k) \wedge o \models lf_i \\ k+1 & \text{otherwise} \end{cases}$$

for any object $o \in \mathcal{O}$. The ordering of constraints gives further information about the ordering of relevance where each object belongs to the highest priority logical formula that is satisfied by the properties of the object. For fuzzy constraints, we assume the existence of fuzzy logical operators and functions that merge sorted lists containing objects and scores.

To further enhance the functionality of the system, we develop clustering methods to find the common properties of objects for an activity. The aim is to help the user by showing relevant properties of objects for an activity beyond those that are specified by the user. Being able to identify and sort files in

relationship to an activity and find the most relevant properties of objects for an activity allows us to perform the following set of tasks on top of the enhanced search queries that we discussed earlier:

- *Show me the files on the visit to Company Acme last year.* Find the dates, people involved in the visit, files created for the trip and organize them in the order of relevance together with the relevant categories of information.
- *Organize my emails based on the known activities.* Parse important properties for each activity and place each mail in one or more activities based on how well they match the given activity (how many properties it matches).
- *Limit my keyword search to those items relevant to activity "Writing the activity paper".* Order the matching items with respect to their match to the given activity.
- *Hide all items relevant to activity "Car Purchase" in all my searches.* Given a level of sensitivity, do not show the items that appear to be related to a specific activity. For example, in a professional setting, do not show files related to personal use of the same computer. This allows the user to implement their own notion of privacy in different settings.
- *Order all files based on their relationship to this file.* Given a video clip, we can find other related items such as presentations we have given with that video clip or the people we met during these meetings. We can also limit the search to a specific activity to focus the search further.
- *Show me all related activities for a specific time/person/place.* If a number of activities are known to the computer, than we can search and find out which activities we were involved in a specific period of time or a given place. This allows us to recall "history" as it is relevant to us.

We are currently working on a prototype of our system to illustrate the above mentioned functionality.

3 Related Work

When the available information is stored on the users' desktops, it is important for information management applications to be able to model users' interpretation of their data and to capture the possibly different meanings, semantics links, and relationships that the users associate to the information units available. For this purpose, various Personal Information Management tools are being developed to assist the user with her navigation/browsing over various forms of personal digital data [10, 5, 4, 8, 13, 12].

MyLifeBits [10] is a research project and a software environment which aims at storing, in digital form, *everything* related to the activities of an individual and providing full-text search, text and media annotations, and hyperlinks to personal data. Another Microsoft project, *Stuff I've Seen* [5], aims at managing personal data, such as already-read email messages, for reuse. Retrieval and presentation of information are based on contextual cues, such as time and author in the case of email.

Recently, there is more work on personal desktop information management. Chandler [4], for instance, is an interesting open source example of such management tools, integrating calendar, email, contact management, task management, notes, and instant messaging functions. Haystack [8] and Gnowsis [13, 12] are systems that adopt the semantic web data modeling approach, and treat all the data objects stored on the desktop as resources on which semantic networks are defined using the Web Consortium's *Resource Description Framework (RDF)* [11].

More user centered treatment of object semantics recently lead to a new emerging research area referred to as *Experiential Computing* [6, 2, 1]. According to this approach, the user interaction systems should exploit and reflect as closely as possible users' previous experiences. Thus, users should be part of the complete system. Experiential environments allow a user to directly observe data and information of interest related to an event and to interact with the data based on his or her own interests in the context of that event. By developing experiential environments, researchers aim to develop new generation information management systems which transform database applications from being simply information sources to being powerful insight and experience sources. The data generated for each event is experienced by an observer and interpreted to create knowledge. In this knowledge production process, the observer plays an important role to interpret the data, and capture the experienced semantics. Recently, there is interest in developing methods to exploit relationships between objects for data cleaning problems [7].

Our approach differentiates from all of the above systems. Based on the fact that objects in a desktop may be related to each other in different ways in different contexts, we argue that users create and modify data as a function of activities that they are involved in. The relatedness of an object to an activity is a fuzzy notion. We develop methods to define and query activities. This allows users to not only locate relevant information but also organize their desktop in relationship to these activities.

4 Conclusions and Future Work

Our notion of an activity - a way to group objects in a user's desktop into overlapping clusters of related objects and related properties - is a first step towards solving the problem of scale when dealing with an ever increasing amount of data both on our own desktop as well as in other data sources that we use and share. Even though available semantic information such as free text or semantic annotations can be consumed easily in any desktop system including ours, generating this information is still very resource intensive. Similarly, content-based retrieval methods for image, video and other media suffer from the problem of being too general. The content of an image may be described very differently based on context. Hence, there is a need to integrate these methods with other data organization methods such as activities to facilitate their effective use.

We are in the process of implementing our prototype activity search and browse system as described in this paper. To this end, we are investigating various algorithmic and system issues in the implementation of this system. One of the main future problems we need to address is the issue of structured activities where an activity may be described by combining simpler activities. An activity may have many different aspects, for example a trip has a preparation phase, the actual trip followed by the other related activities. Based on our queries, we might be interested in a certain aspect of a given activity and the system should immediately adapt to this using a form of relevance feedback. Even though we can keep activity definitions fairly simple, we can learn about user's specific preferences based on their interactions with the system and integrate these back into the system. Our long term goal is to augment the desktop with inference tools that make use of the semantic data available in the activities to automatically associate semantics with data objects. The availability of these solutions would be an important first step towards solving the problem of scale in information systems.

Acknowledgment. We would like to thank Ramesh Jain for stimulating discussions on multimedia querying and experiential computing.

References

1. P Appan, H. Sundaram, D. Birchfield, "Communicating everyday experiences" *Proceedings of the 1st ACM workshop on Story representation, mechanism and context*, 2004.
2. S. Boll , U. Westermann, "Mediaether: an event space for context-aware multimedia experiences", *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*, 2003.
3. Jan Chomicki: Preference formulas in relational queries. ACM Trans. Database Syst. 28(4): 427-466 (2003).
4. "Vision of Chandler", *www.osafoundation.org*, 2005
5. S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. "Stuff i've seen: A system for personal information retrieval and re-use." *Proceedings of SIGIR*, 2003.
6. R. Jain. "Experiential computing", Commun. ACM, vol.46(7), 2003, pp. 48–55.
7. D. V. Kalashnikov, S. Mehrotra, Z. Chen: "Exploiting Relationships for Domain-Independent Data Cleaning." SDM 2005.
8. D.R. Karger, K. Bakshi, D. Huynh, D. Quan, V. Sinha: "Haystack: A General Purpose Information Management Tool for End Users of Semistructured Data." *Proc. CIDR 2005.*
9. F. Manola and E. Miller: "RDF primer". *www.w3.org/TR/rdf-primer/*, 2003.
10. "MyLifeBits Project", *research.microsoft.com/barc/mediapresence/MyLifeBits.aspx*, 2005.
11. "Resource Description Framework (RDF)" *//www.w3.org/RDF/*, 2005.
12. L. Sauermann: "The Semantic Desktop - a basis for Personal Knowledge Management." *Proc. I-KNOW 05.*
13. L. Sauermann: "The Gnowsis Semantic Desktop for Information Integration" *Proceedings of IOA Workshop of the WM2005 Conference.*