Building a Semantic Representation for Personal Information

Jinyoung Kim, Anton Bakalov, David A. Smith and W. Bruce Croft
Department of Computer Science
University of Massachusetts Amherst
{jykim,abakalov,dasmith,croft}@cs.umass.edu

ABSTRACT

A typical collection of personal information contains many documents and mentions many concepts (e.g., person names, events, etc.). In this environment, associative browsing between these concepts and documents can be useful as a complement for search. Previous approaches in the area of semantic desktops aimed at addressing this task. However, they were not practical because they require tedious manual annotation by the user.

In this work, we suggest a methodology and a prototype system for building a semantic representation of personal information based on click feedback from the user. We employed a feature-based model of associations between the concepts and documents. Our initial evaluation shows that the suggested semantic representation can play an important role in the known-item finding task and that the system can learn to predict such associations with a small amount of click data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: [Information Search and Retrieval]

General Terms

Algorithms

Keywords

Personal Information Management, Semantic Desktop, Associative Browsing

1. INTRODUCTION

Although considerable effort has been made to find a more effective solution to personal information management (PIM), it seems that we have not made much progress for the last 40 years. Personal information on desktops is still organized in files and folders, often making it impossible to recall where it is stored. Recent web services that manage a subset of our personal information (e.g. emails, calendar items, bookmarks) are adding more confusion by each having their own information organization.

Having an effective search system can alleviate this problem to some extent, yet it has several limitations. First of all, finding the right keywords for searching is difficult in many cases. More importantly, searching addresses only a part of the whole problem because exploring or organizing one's own information is also an important part of PIM activities [8]. Some studies [13] [1] suggest that people want to browse even if they have an effective search system at their disposal.

Researchers tried to overcome these limitations by building an additional structure of personal information [9] [6]. Such techniques are generally referred to as a 'semantic desktop' - the organization of personal information across devices and applications based on concepts and the relationships between them. With a semantic desktop, instead of navigating through files and folders, users can browse their personal information based on people's names, events and the relationship between them. Given that people remember things largely by associations [2], this model of interaction is likely to considerably enhance the accessibility of personal information.

Despite this appealing vision, none of the approaches suggested previously have been widely adopted. According to a recent user study [12], the most conspicuous problem is that these systems have a complex data model and interface, making them hard to understand and maintain for the end user. A related issue is that users need to make manual annotations (e.g. Tom is-a-friend-of Mary) to populate the data model. Overall, the additional structures in these systems seem to have come with a cost users are not willing to pay.

In an effort to keep the benefit of structured data while minimizing the cost for the user, we propose a simple model of semantic representation for personal information. It is composed of (information) *items*, *tags* and the *links* between items. Items here represent information objects with textual contents. These objects can be *documents* collected from many sources, or *concepts* – entities and terms of interest to the user. Tags and links are the metadata that enables the grouping or the association of individual items.

This model eliminates many of the complications that existing approaches have. Firstly, the fundamental unit of managing information is documents and concepts that can be tagged. We believe that this may be more intuitive than

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

the RDF-based ontologies employed in most previous work. Secondly, although the system internally maintains many types of links between items, users are presented with a simple ranked list of related items. This approach is found to be sufficient for providing an effective browsing capability [12].

More importantly, this simplified data model makes it possible to automatically build and maintain the structure to a great extent. When documents are imported into the system, concepts, tags and different types of links can be extracted from their metadata. Given this initial set-up, users can start searching or browsing their own information. Meanwhile, the system collects implicit feedback from these activities to enrich its data model and fine-tune parameters.

The rest of this paper is organized as follows. In the next section, we provide an overview of related work. Our data model and the prototype system for building a semantic representation of personal information are introduced. We then describe the learning method to create suggestions for associative browsing, followed by the evaluation of the semantic representation and the learning approach we employed.

2. RELATED WORK

The term 'semantic desktop' encompasses several other methods for building a semantic representation of personal information. Sauermann et al. [11] provides a good overview of the research efforts up to 2005. Among these approaches is the IRIS semantic desktop [4] that provides a platform from which users can manage all of their information. In the small-scale evaluation mentioned in the paper, however, they found that the user interface was not sufficiently fast and robust for real-world use. Recently, Sauermann et al. [12] built and evaluated the Gnowsis semantic desktop and found that users in general are not willing to make annotations to their database except for simple tagging. While these projects started with a grand vision, complex data models based on ontologies as well as the need for manual data entry seems to have prevented them from having a wide adoption.

More recently, Chen et al. [3] proposed a system called iMecho. It builds a link structure between documents primarily based on user activities. This structure is used to compute a query-independent authority score in a PageRank-like manner. Our work is different from iMecho in that we focus on browsing based on similarity search while their system focuses on improving keyword search. Also, we utilize the user's feedback to learn personalized feature weights automatically, whereas they let the user adjust weights manually. Furthermore, while they employed a faceted browsing model that extracts facets from a fraction of the document metadata, our notion of concept space is more general and comprehensive.

3. DATA MODEL & ARCHITECTURE

In this section, we first introduce our data model. Then, we describe the prototype system implementing the data model.

3.1 Data Model

As briefly discussed in the introduction, our goal in designing the data model is minimizing the complexity while preserving the benefit of rich semantic representation. In

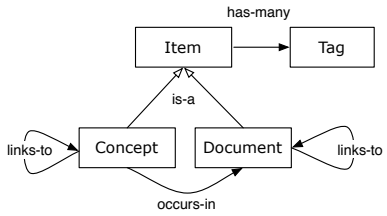

Figure 1: Suggested data model for personal information.

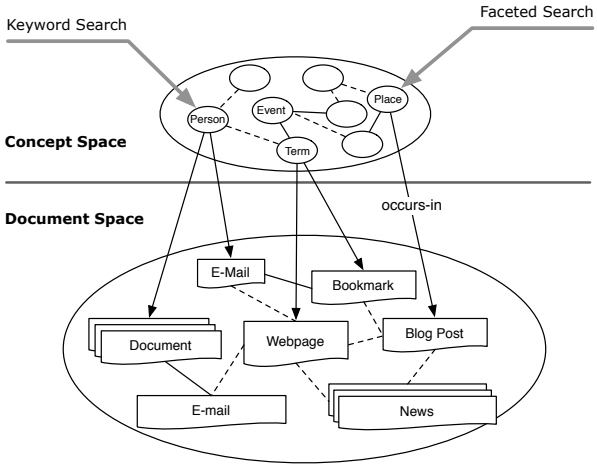

Figure 2: An example of link structure between items. You can start browsing from any item by keyword or faceted search.

what follows, we describe each component of the data model in greater detail.

3.1.1 Items - Documents and Concepts

Our definition of semantic representation consists of the collection of *items* that can have *tags* and *links*, as seen in Figure 1. Since items can have a text representation – title, URI, content and metadata – the user can perform keyword search. Tags and links allow the user to browse the document collection.

As for the subtypes of items, one can see that all documents are items. Another important item type is the concept, which can be entities and terms of interest to the user. Although it is tempting to treat concepts as a different class of objects from documents, we consider them as items because they shares many of the features that other items have – they can be tagged, linked or have a text representation (e.g., homepage of a person or a wikipedia entry). Concepts are treated differently from documents, however, in that the system extracts concepts from documents and builds a separate space that one can use to browse one's information. Figure 2 shows a diagram of the document and concept space.

3.1.2 Link Structure between Items

An important feature of our semantic representation is the link structure between items, which enables associative browsing. This is where previous solutions fell short since many of them required the user to create the structure manually. Some researchers suggested automatically extracting links [6] [5] but their methods were applicable only in some special cases (e.g. an email and its attachments).

Our data model is similar in that we have the links of many types extracted from different sources. However, the way the link structure is presented to the user makes our approach different from what was described above. Instead of presenting many types of labeled relationships between items separately, we present the user with a rank list of related items, generated by combining scores for each link types using appropriate weights. This simplification comes with a cost – the user can no longer see the type of relationship between items – but it was found [12] that the nature of the relationship is not important for purpose of associative browsing.

More importantly, this unified notion of semantic similarity enables the system to create a single semantic structure captured by the links shown in Figure 2. Furthermore, since the characteristics of the collection and the notion of semantic similarity might vary significantly for each user, we can adjust the weight of each link type appropriately using the click feedback from the user.

3.2 System Architecture

In this section, we introduce the prototype system we developed for this project - LiFiDeA[1]. First, it collects documents and extracts concepts from document metadata. After this initialization step, the user can browse the space of concepts and documents. At the same time the system saves click feedback and uses it for learning the feature weights.

3.2.1 Document Collection

First, LiFiDeA collects personal documents from various sources including desktop files, email and websites that provide RSS feed. This functionality allows the system to work with all of users' personal information on computing clouds as well as desktop because most web services offer an RSS feed. During this collection step, the system also extracts metadata (e.g., authors and recipients of emails, tags of blog postings) associated with each document type.

3.2.2 Concept Creation

Given a collection of documents, the next step is to create concepts (e.g., names, domain terms, and so on), which constitute an extra layer one can use to browse documents. In LiFiDeA, concepts are another *items* like documents, as described in Section 3.1.1. However, they are different in that the occurrences of concepts are extracted from documents.

In addition to automatically extracting concepts from documents (e.g., authors of emails), the systems allows the user to create concepts. There are several ways of adding concepts in LiFiDeA. A user can choose to promote a tag to a concept. Secondly, the user can decide to convert an appropriate document (e.g. Wikipedia article) to a concept. It is also possible to create a concept out of query words that the user types in to find documents.

3.2.3 User Interface for Browsing and Searching

The web interface shown in Figure 3 allows the user to freely browse the concept and document space. In the back, you can see a index page showing the list of publications along with tags. Here, users can perform full-text search by typing in keywords or faceted search by specifying conditions of filtering, which provides a starting point of browsing as depicted in Figure 2.

The front page of a concept 'Search Engine' shows related documents and concepts, created by combining scores from each link types as features. When the user clicks on this ranked list, the system collects the user's clicks on relevant concepts and uses them as training data for adjusting feature weights. After each training cycle, the list of relevant concepts improves due to more refined feature weights.

[1]We coined this name by combining the words 'life' and 'idea'.

4. LEARNING FRAMEWORK

One of the core components of our system is the rank list of related concepts or documents, generated by combining features with appropriate weights. In this section, we briefly describe features and learning methods we used.

As for features, we used well-known measures of textual similarity such as string edit distance and cosine similarity. Since each item has a timestamp associated with it, we used the temporal proximity as another feature. In addition, tightly interconnected spaces of documents and concepts gave another opportunities for computing the similarity between items. For concepts, we could use the contents of related documents to measure the similarity. For documents, we can use the sets of connected concepts.

In learning the weight of each feature, we used two algorithms – iterative grid search and RankSVM [7] with different characteristics. In terms of the objective function, grid search simply finds the set of parameters that maximizes the target metric, whereas the goal of RankSVM is to predict the pairwise preference relation with highest accuracy. Also, while grid search uses each click as a relevance judgment, RankSVM interprets each click as a pairwise preference.

5. EVALUATION

In this section, we briefly describe the methods and results of evaluating our approach. We did two rounds of game-style user studies in which participants were asked to perform a given task in a competitive environment. In the first round, users were asked to find the target document using only search and associative browsing between documents. In the second round, concepts were available for searching and browsing, thereby providing a full access to the semantic representation.

Our evaluation suggests that the semantic representation is useful for known-item finding tasks, especially when concepts are used in addition to documents. Users browsed a lot, and that led to higher success rates. We also found that the combination of features significantly improves the quality of suggestions, and that learning combination weights takes less than 100 clicks. More details can be found in our technical report [10].

6. CONCLUSIONS

In this paper, we suggested a data model and a system for building a semantic representation of personal information. Our semantic representation is composed of items (concepts and documents) that can be tagged and the links between them. Instead of displaying links of many types as they are, we generate a single ranked list of related items by combining the scores of many link types with appropriate weights, where weights are learned using click feedback from the user.

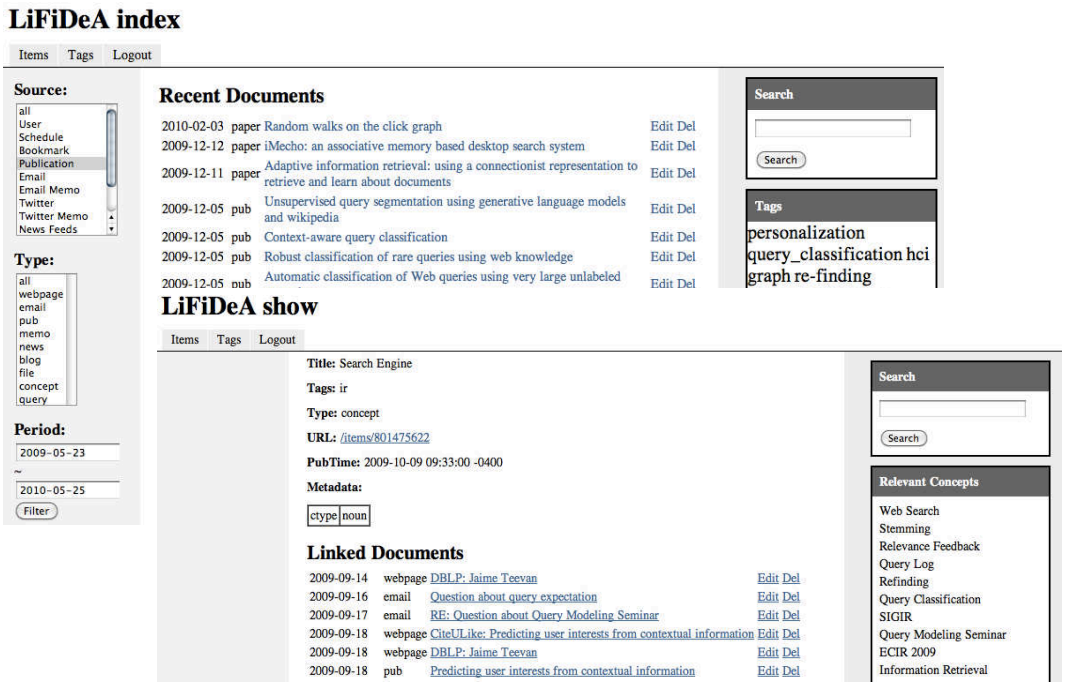

Figure 3: LiFiDeA user interface. Back: Index page showing the list of publications along with tags. Front: The page of a concept 'Search Engine' showing related documents and concepts.

In our recent evaluation based on a series of user studies, we found that the semantic representation is useful for known-item finding tasks, and that the system can learn to predict such associations with a small amount of click data. As a future work, we plan to perform a large-scale evaluation both in a real user environment and in the context of a specific task.

7. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0707801, and in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved search engines and navigation preference in personal information management. *ACM Trans. Inf. Syst.*, 26(4):1–24, 2008.

[2] D. H. Chau, B. Myers, and A. Faulring. What to do when search fails: finding information by association. In *CHI '08*, pages 999–1008, New York, NY, USA, 2008. ACM.

[3] J. Chen, H. Guo, W. Wu, and W. Wang. imecho: an associative memory based desktop search system. In *CIKM '09*, pages 731–740, New York, NY, USA, 2009. ACM.

[4] A. Cheyer, J. Park, and R. Giuli. Iris: Integrate. relate. infer. share. *1st Workshop on The Semantic Desktop. ISWC '05*, nov 2005.

[5] P.-A. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle^{++}: Semantically enhanced searching and ranking on the desktop. In *ESWC*, pages 348–362, 2006.

[6] X. Dong and A. Y. Halevy. A platform for personal information management and integration. In *CIDR*, pages 119–130, 2005.

[7] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02*, pages 133–142, New York, NY, USA, 2002. ACM.

[8] W. Jones and J. Teevan. *Personal Information Management*. University of Washington Press, 2008.

[9] D. R. Karger, K. Bakshi, D. Huynh, D. Quan, and V. Sinha. Haystack: A general-purpose information management tool for end users based on semistructured data. In *CIDR*, pages 13–26, 2005.

[10] J. Kim, A. Bakalov, D. A. Smith, and W. B. Croft. Evaluating a semantic representation for personal information. CIIR Technical Report. Univ. of Mass., Amherst, 2010.

[11] L. Sauermann, A. Bernardi, and A. Dengel. Overview and outlook on the semantic desktop. In Dennis and L. Sauermann, editors, *ISWC '2005*, 2005.

[12] L. Sauermann and D. Heim. Evaluating long-term use of the gnowsis semantic desktop for pim. In *ISWC '08*, pages 467–482, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04*, pages 415–422, New York, NY, USA, 2004. ACM.