Semantically Annotating the Desktop

Towards a Personal Ontology

Jimmy Borg
Department of Computer Science and AI
University of Malta
jbor059@um.edu.mt

Matthew Montebello
Department of Computer Science and AI
University of Malta
matthew.montebello@um.edu.mt

ABSTRACT

The advent of the World-Wide Web brought with it a proliferation of information from e-mail, forums, chat, sites that rapidly led to information overload and a subsequent storage problem and maintenance on users' personal computers. The desktop has become a repository of data that hosts various types of files. The recent massive increase in data has resulted in a continuous attempt to enhance our data organisation techniques and hence to the development of personal information management software.

In this paper we present an overview of data organisation techniques related to personal data management that have been an active research area for decades. We will look at how personal information managers handle different types of files, and abstract these file types into a single user interface. Despite their advanced user interfaces, we argue that traditional personal information managers tend to be very domain specific and lack in user adaptability. To address these limitations we propose a semantic desktop application that exploits the flexibility of semantic web technologies, and introduces the concept of a Personal Ontology to aid in data organisation and can be used by other desktop applications such as information retrieval and intelligent software agents.

1. INTRODUCTION

With the introduction of the PC, computers have moved from single purpose to multipurpose machines. Personal Computers are no longer only used to maintain a database or to run a payroll application. PCs have become an integral part of our daily life. On our PC we watch videos, play audio files, watch TV and store collections of music CDs that we used to place on a shelf. We can take, store and share digital photos. We can chat, write emails, maintain calendars and reminders and store our contact information list. Nowadays books are being stored in electronic format, libraries are becoming online bookstores, magazines and newspapers are being published digitally and huge collections of scientific papers and articles are accessible through the World Wide Web.

This popularity of the PC and the World Wide Web has exposed our machines to a huge amount of new data that needs to be stored, maintained and easily accessed. It is no longer a question of whether we have the information, it has become a question of how we are going to find the required information. Web search engines do the job very well but unfortunately their desktop counterparts are still quite limited. Documents on the desktop are not linked like web pages and thus algorithms such as PageRank cannot be used [3]. Ironically enough, in some cases one may find it more efficient to search for the information on the Web rather than on his or her own personal computer.

In the rest of the paper we will be discussing the organisation and retrieval of personal information. In Section 2 we will define what we mean by personal information, then we will discuss how we can manage personal information and finally we will give an overview of different types of personal information management software. In Section 3 we will proceed by proposing the Semantic Desktop, a system that semantically annotates the data on a user's personal computer and, by using standard Semantic Web languages for information representation, creates a Personal Ontology. We will conclude the paper by discussing some possible environments where the Personal Ontology can be used.

2. PERSONAL INFORMATION

Personal information can be defined as data that a single user stores on his or her personal computer. This information can be of different types and we can produce a never ending list of information that can be classified as personal. To get an idea, such information might include;

- Calendar Entries such as birthdays, anniversaries, appointments, meetings and other significant dates,
- Email Repositories,
- Instant message archives,
- Contact information such as telephone numbers, mobile numbers and postal and email addresses,
- Files of various types such as documents, papers, photos, digital books, video clips and web pages,

CSAW 2006, MCST Bighi, Malta

- Various types of lists such as reminders, notes, bookmarks and RSS/Atom Feeds.

In the rest of this section we will discuss ways of how this personal information can be organised, in other words, personal information management. We will then give an overview of different applications that aid in personal information management, also known as personal information managers.

2.1 Personal Information Management

The area of personal information management has a long history composed of very interesting examples that helped in shaping today's theories. Some even date back to the pre-computer area, such as the famous article "As we may think" [2] by Dr.V. Bush. In his article of 1945, Bush describes his visionary system, *Memex*, as

> *"a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."*

Although limited to only analog devices, *Memex* not only was able to store different types of mediums but it allowed searching to be performed in an associative matter, like the human mind. We can say that today we are very close to this kind of system since the computer can store and access many different types of files. Although memory is now becoming much more affordable, it is not feasible and not necessary for all the data to be stored on the same machine. Nowadays the web can be seen as a repository of data that extends our own storage space. Recent approaches to information retrieval and organisation such as [1] are based on this idea.

Despite the fact that our mind thinks in an associative manner, we cannot simply eliminate the traditional indexing and categorisation approach. This approach is the most widely used and over time it has proved itself to be very efficient and effective in many situations. For example one cannot question an indexing approach on a telephone directory or a categorisation approach on a web directory. However we might question other areas that we may take as obvious, such as, "Is a tree organisation the most suitable for a file structure?" and "Is a relational database an appropriate storage method for an email repository?". We will discuss these issues in more detail in Section 3. We will now proceed by looking at some personal information managers that are widely used, and highlight similarities and possible improvements to these tools.

2.2 Personal Information Managers

Personal information managers, or PIMs, are tools that help users to store, maintain stores, search and retrieve personal information. In other words a personal information manager's aim is to aid in the organisation and retrieval of data in a single user perspective. Typical challenges that personal information management software encounter are:

Huge amount of information, As we already discussed in Section 1, personal information managers typically deal with huge amount of information that we use everyday, being either information that we directly access, for example when reading an email, or information that is indirectly accessed by the application, for example when checking for new emails.

of different type and nature, Information does not only consist of different types of files but also databases, archives and online links. For a typical list of personal information that a user may regularly use one can refer to the beginning of this section.

coming from different sources The information may not only be stored on the computer's hard drive but can also reside on network drives, servers, web pages or other online repositories.

As the amount of information that we deal with everyday is increasing, personal information managers are becoming more popular since they can minimise the burden from our memory. One can find applications of different flavours that target different types of users. Some users use a PIM for storage and retrieval of data. Others use it as a communication tool, to send emails, fax and instant massages. Others try to make their lives more organised by keeping important calendar dates, to-do lists and meeting reminders. In general we can categorise information management software as PC based, web based or PDA.

PC based packages are the oldest and typically tend to be the most feature oriented. Most consist of email programs, contact list, organisers and maybe a calendar. A typical and very widely used application is Microsoft Outlook [17], which packages many features under a single user interface. Other applications such as Lotus Notes [18] offer a networked flavour, typically more oriented for the business class. The application also includes an advanced semi-automatic meeting scheduler.

Web based solutions usually take an organisational approach and unlike the PC based applications lack the storage of large data. Typically these applications range from email clients to calendars and schedulers. Web applications offer the advantage of accessibility from any internet enabled machine, not only from the user's personal computer. A typical example of a web based system is the relatively new Google services, which range from an email client, calendar, scheduler and a document editor.

Personal Digital Assistants, or PDAs, are mobile devices designed for being used as personal organisers. Their main merit is mobility, on the other hand, they usually lack in memory and their functionality depends very much on the operator's connectivity. The functionality of PDA software is very similar to web based systems and typically includes a combination of email, calendar, reminder, address book and notes. An interesting, typical feature of PIM software on a PDA is that it can synchronise with other PIM software on a personal computer.

In general, these systems are targeted at different classes of users. The information structure and functions are built upon the targeted user's needs, for example Microsoft Outlook is targeted for a typical home user that needs to access

mail and maybe keep a simple calendar of events. On the other hand Lotus Notes provides features that are more targeted for the business class of users, providing them with a more advanced meeting scheduler, email access over a networked environment and an advanced user profile system. In the next section we will argue that by focusing on the meaning of the data, rather than the user we can build an application that adapts itself according to the user's needs.

3. THE SEMANTIC APPROACH

> *"The dream behind the Web is of a common information space in which we communicate by sharing information. Its universality is essential: the fact that a hypertext link can point to anything, be it personal, local or global, be it draft or highly polished. There was a second part of the dream, too, dependent on the Web being so generally used that it became a realistic mirror (or in fact the primary embodiment) of the ways in which we work and play and socialize. That was that once the state of our interactions was on line, we could then use computers to help us analyse it, make sense of what we are doing, where we individually fit in, and how we can better work together."* -Tim Berners-Lee [4]

That was the initial vision of Tim Berners-Lee, the creator of the World Wide Web as we know it today. However we can note that the second part of his dream is not yet achieved, and this is where we are moving to, the Semantic Web. He defines the Semantic Web as

> *"an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* -Tim Berners-Lee [5]

In the rest of this section we will discuss how we can apply information management techniques and Semantic Web technologies to the personal computer in order to improve personal information management and collaboration.

3.1 A Semantic Desktop

Semantic desktop applications are quite innovative. The idea evolved from the vision of the Semantic Web itself. The semantic annotation of data on the desktop will allow for the integration of desktop applications with the Semantic Web. The personal computer, as a repository of data, can be seen as a small Web in itself. By annotating the data on the PC we will be placing the first building stone for the Semantic Web. Since the Semantic Web is still in its infancy, current semantic desktop applications are generally built for research purposes. Some of these applications include:
Haystack system at MIT [6],
Gnowsis system at DFKI [10],
D-BIN by SEMEDIA [11],
OpenIris by SRI [12] and
Chandler system by the OSA foundation [13].

A typical semantic desktop application can be split into three main components, namely, the ontology, applications that create and maintain the ontology and applications that make use of the ontology. The foundation of the system is the ontology. This uses standard general purpose languages for information representation, such as RDF and OWL [14]. A typical ontology stores the semantic metadata of all the personal information of the user, thus in this paper we will refer to it as a *Personal Ontology*. The Personal Ontology consists of three levels. At its most basic state, the ontology describes the structure of basic file types. We can call this the *Storage Ontology*. The *Preferences Ontology* is used to store the user's preferences and settings, which are not only used by the semantic desktop application but can be used by all other applications. The third level of the *Personal Ontology* is the *Content Ontology* and is used to describe the content of the files.

The second component of a Semantic Desktop system is a mechanism that modifies the underlying ontology. This will typically consist of several function specific applications, such as an email client, a file browser and a calendar. There are two approaches that one can take when developing a semantic application; the monopolistic approach and the integrative approach. In a monopolistic approach, the semantic desktop application will replace many existing applications and group all the functions into a single user interface. Although this method generally sounds neater, it requires the user to adapt to a new system. An example of such system is Haystack. The integrative approach, adapted by the Gnowsis application, will add extra functionality into the existing applications and make them interact with the ontology. This does not require the user to learn a new system and it can reduce the development effort. However such approach may be limited by the flexibility of the third party applications.

The ontology can be virtually composed of any type of personal information. It is the applications that create and maintain the ontology that limits the extent of the Personal Ontology. Different systems provide different sets of applications, depending on their scope and size of the project. Typical, basic functions that one will find present in almost all systems are emails, calendar events and browser cache. A common challenge in these systems is the annotation of the file system. The user spends considerable time building complex folder classification hierarchies thus it is of vital importance for the semantic desktop application to use this information. However the current operating systems lack the much needed support for file-handling event triggering. [3] proposes a similar application that uses an in-notify enabled Linux kernel while [9] proposes a similar system on Windows. Having the Personal Ontology created and appropriately maintained, it will become a question of how the ontology can be used.

3.2 Using the *Personal Ontology*

We will proceed by identifying possible ways of utilising the Personal Ontology, most of which consist of quite novel research areas. Data organisation is the most obvious utilisation of the ontology, and is the main subject that we have discussed till now. The difference between a semantic desktop application and other personal information managers is that the semantic organiser can change the way of presenting the information to the user according to the information

itself. We can illustrate this by a simple scenario regarding the usage of contact information; in a company the manager will need to know detailed information about a contact such as the name and surname, telephone, fax and mobile numbers and email and postal addresses. On the other hand in personal contact list used only for telephone numbers the user might need to store the name and surname, or maybe a nickname, the telephone number and possibly a mobile number. As discussed in [7], by building the user interface upon the Storage Ontology, the user will be presented with only the required data.

A key element in every user adaptive system is the context information. The Personal Ontology can be an invaluable element for making an application user adaptive. Since the ontology uses standard semantic web languages, it can be accessed by any semantic web application, not just by the Semantic Desktop system. An approach that is quite new to desktop applications is the use of content ontology, partially described in [8]. The idea behind the Content Ontology is to semantically annotate the content of the files, especially documents and emails. The application will then be able to analyse the Content Ontology of different files and suggest possible relations between the files. While the Content Ontology can make the system more adaptive, the Preferences Ontology can make the system more adaptable and share a generalised set of user preferences between several applications.

As P.A. Chirita et al states, in [3], current approaches to desktop search, such as Google Desktop search [15] on Windows or Beagle [16] on Linux, do not include metadata in their system but only perform searching using regular text indexing. This causes such systems to perform poorly when compared to their web counterparts. The key element that makes web search systems very effective is the linking between the elements, which is virtually inexistent on current file systems. The Personal Ontology, especially if the Content Ontology level is implemented efficiently, could fill this gap. The document links can help to apply result ranking techniques [19] in desktop search algorithms.

The Social Semantic Desktop can be seen as a networked collection of Semantic Desktops. The idea is to create an environment where data and metadata can be easily shared between peers. Peers, or agents on personal computers, can collaborate together and form communities to exchange knowledge while reducing the time for users to filter and file the information [20]. The Semantic Desktop is one of the three main components of the Social Semantic Desktop. The Semantic Desktop system, in conjunction with Peer To Peer services, provides a mechanism for users to share their information. The third component, Social Software, maps the social connections between different people into the technical infrastructure.

Other systems use the ontology for more specific purposes. For example IRIS provides a Semantic Desktop interface that builds a desktop ontology which will be used as a learning environment for the CALO Cognitive Assistant project [22]. CALO, [21], is a personal assistant that learns by applying automated machine learning techniques on a user's personal data.

4. CONCLUSION

In this paper we have presented techniques that can be used to semantically annotate personal information on a user's personal computer, thus creating a Semantic Desktop. Having the data semantically annotated using standard Semantic Web languages make it possible for applications to integrate the desktop with the Semantic Web.

To conclude, we can say that the Semantic Desktop system goes beyond the purpose of data organisation. The Personal Ontology can be adopted and used for different purposes ranging from file organisation, to machine learning environments, to the creation of a large semantic network where both users and applications can reason about the shared knowledge.

5. REFERENCES

[1] D. Elsweiler and I. Ruthven and L. Ma, *Considering Human Memory in PIM*, SIGIR 2006 Workshop on Personal Information Management, August 10-11, 2006, Seattle, Washington

[2] Vannevar Bush, *As We May Think*, The Atlantic Monthly, 176(1), p101-108, July 1945

[3] P.A. Chirita and R. Gavriloaie and S. Ghita, *Activity Based Metadata for Semantic Desktop Search*, In Proc. of Second European Semantic Web Conference, ESWC2005, May 21 - June 1, 2005, Heraklion, Crete, Greece

[4] Tim Berners-Lee, *The World Wide Web: A very short personal history*, http://www.w3.org/People/Berners-Lee/ShortHistory.html

[5] Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

[6] Haystack Project, http://haystack.lcs.mit.edu/

[7] E. Adar and D. Karger and L.Stein, *Haystack: Par-User Information Environments*, Conference on Information and Knowledge Management, 1999

[8] B. Katz and J. Lin and D. Quan, *Natural Language Annotations for the semantic web*, ODBASE, 2002

[9] Leopold Sauermann, *The Gnowsis, Using Semantic Web Technologies to build a Semantic Desktop*, Master's Thesis, TU Vienna, 2003

[10] Gnowsis Project, http://www.gnowsis.org

[11] D-Bin Project, http://www.dbin.org/

[12] IRIS Semantic Desktop Project, http://www.openiris.org/

[13] Chandler Project, http://chandler.osafoundation.org/

[14] Web Ontology Working Group

[15] Google Desktop Search Application, http://desktop.google.com/

[16] Gnome Beagle Desktop Search, http://www.gnome.org/projects/beagle/

[17] Microsoft Outlook, www.microsoft.com/outlook/

[18] Lotus Notes, http://www.lotus.com/notes/

[19] Stefania Costache, *Using Your Desktop as Personal Digital Library*, TCDL Bulletin, 2006

[20] S. Decker and M. Frank, *The Social Semantic Desktop*, DERI Technical Report 2004-05-02, May 2004

[21] CALO Project, http://www.ai.sri.com/software/CALO

[22] A. Cheyer and J. Park and R. Giuli, *IRIS: Integrate. Relate. Infer. Share.*, In Proc. of Fourth Intl. Semantic Web Conference Workshop on the Semantic Desktop, Galway, Ireland, Nov. 2005