Using Superimposed and Context Information to Find and Re-find Sub-documents

Sudarshan Murthy[1], Uma Murthy[2], Edward A. Fox[2]

[1]Department of Computer Science, Portland State University, PO Box 751, Portland OR 97207

[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060

smurthy@cs.pdx.edu, {umurthy, fox}@vt.edu

ABSTRACT

Superimposed information (SI) is new information such as annotations created in reference to bits of existing information. SI references existing information using an abstraction called *mark*. A mark can be *activated* to see referenced information in its original context. *Context information* is the set of information obtained using a mark. For example, page number and font characteristics are included in the context information for a selection in a PDF document.

Superimposed and context information can make finding and re-finding of sub-documents easier than is possible with the state of the art. For example, search results can embed marks which users can use activate to navigate directly to sub-documents; search results can include context information to potentially reduce the number of click-through operations needed to locate relevant information.

General Terms

Management, Design, Human Factors

Keywords

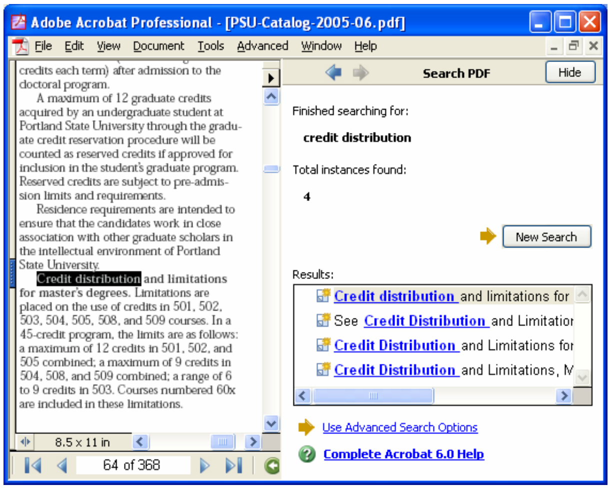
Personal Information Management, Superimposed Information, Context, Finding, Re-finding

1. INTRODUCTION

Consider this scenario: A student wishes to *find* information on 'credit distribution' in her university's catalog. She first uses a document search function on her university's web site to locate the catalog. She then uses the search function in Adobe® Acrobat® to locate the phrase 'credit distribution' in the PDF version of the catalog. Figure 1 shows the result of this search performed in Acrobat Version 6. The search function lists the four occurrences of the search phrase in the document. It also provides some *context information* for each occurrence, in the form of some text that surrounds the search phrase as it appears in the document. Hovering the mouse cursor over a result displays a page number where the phrase occurs. Clicking on a result takes the student to the corresponding occurrence of the phrase in the document (that is, to a sub-document), and the search phrase is highlighted.

The context information, specifically the surrounding text, displayed for each sub-document in the result can enable the student to determine if a particular sub-document might be relevant without explicitly visiting the sub-document. That is, context information can help reduce the number of click-through operations. However, the utility of context information depends on the quality and quantity of information displayed. For example, Adobe Acrobat displays up to 12 words after the search phrase, and sometimes includes one word before the search phrase in the result list. This amount of information may not always suffice to determine if a sub-document is worth visiting. The number of click-through operations may be reduced if the student can selectively see more (or different) context information for a sub-document *without* visiting it.

Note that the student uses *two* distinct search functions to find the information she is looking for: A document search function on the web, and a sub-document search function within Acrobat. Ironically, she uses the same search phrase in both searches, but the document search function (in this case, Google-based site search) returns only documents, not sub-documents.



credits each term) after admission to the doctoral program.

A maximum of 12 graduate credits acquired by an undergraduate student at Portland State University through the graduate credit reservation procedure will be counted as reserved credits if approved for inclusion in the student's graduate program. Reserved credits are subject to pre-admission limits and requirements.

Residence requirements are intended to ensure that the candidates work in close association with other graduate scholars in the intellectual environment of Portland State University.

Credit distribution and limitations for master's degrees. Limitations are placed on the use of credits in 501, 502, 503, 504, 505, 508, and 509 courses. In a 45-credit program, the limits are as follows: a maximum of 12 credits in 501, 502, and 505 combined; a maximum of 9 credits in 504, 508, and 509 combined; a range of 6 to 9 credits in 503. Courses numbered 60x are included in these limitations.

Finished searching for:

credit distribution

Total instances found:

4

Results:

- Credit distribution and limitations for
- See Credit Distribution and Limitatior
- Credit Distribution and Limitations for
- Credit Distribution and Limitations, M

Use Advanced Search Options

Complete Acrobat 6.0 Help

Figure 1: Search results in Adobe Acrobat.

Now, imagine that the student wishes to *re-find* the information on credit distribution. The student first has to recall the location of the catalog, and then the location of the relevant sub-document within the catalog. Alternatively, she needs to recall the search phrase she had used earlier, and when she receives a list of candidate sub-documents, she needs to recognize the sub-document she had consulted earlier. Essentially, she has to repeat

Invitational Workshop on Personal Information Management (PIM 2006) August 10–11, 2006, Seattle, WA.

the finding process, possibly again visiting some documents and sub-documents that are not relevant.

Some tools do assist in re-finding information by allowing users to record results from past finding attempts, but they tend to be either document-centric or document specific. For example, bookmarks in web browsers are document-centric because users can store references only to documents (as URLs). Adobe Acrobat allows a user to create bookmarks to specific *pages* within a document (not to arbitrary sub-documents), but the bookmarks are document-specific because they are stored within a document. That is, using Acrobat bookmarks requires ownership of the document. (The student in the example scenario does not own the university catalog.)

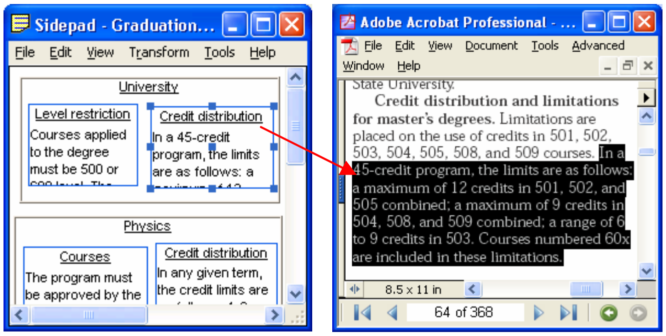
We believe that a single system can assist in finding and re-finding personal information. Ideally, such a system combines document and sub-document searches, and supports references at both complete document and sub-document granularity. It allows references to arbitrary kinds (page, paragraph, line) and sizes of sub-documents, and works with any information format (PDF, HTML, spreadsheets) at any location (local file system, web). It provides context information for documents and sub-documents, and it does not assume ownership of referenced information.

In this paper, we highlight ways our research on *superimposed information* may help build systems with these capabilities. We outline the motivation to use SI in finding and re-finding by tying in related work on personal information management (PIM).

2. SUPERIMPOSED INFORMATION

Superimposed information (SI) refers to new information such as comments, and new structures such as lists, placed over existing *base information* [13]. In this setting, a user creates and manages SI in *superimposed applications* (SAs), which in turn might use existing *base applications* to *activate* (or show in context) sub-documents, and to retrieve context information such as text excerpts and page numbers for sub-documents.

SI uses an abstraction called *mark* [3] to reference sub-documents in documents of any format. This abstraction is defined and implemented in the *Superimposed Pluggable Architecture for Contexts and Excerpts* [17] (SPARCE), our middleware for SI management. We have implemented the mark abstraction for several base document types such as PDF, HTML, Microsoft® Word, and several audio and video formats. Support for new base types can be easily added without affecting existing applications.



Courses applied to the degree must be 500 or

In a 45-credit program, the limits are as follows: a

The program must be approved by the

In any given term, the credit limits are

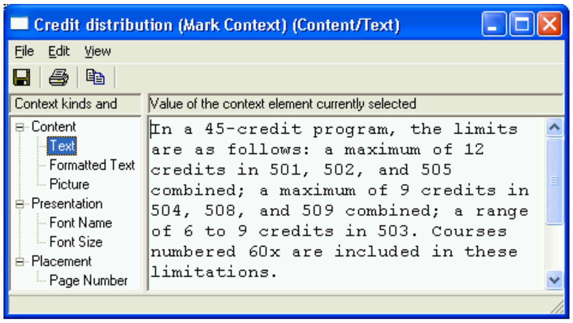
State University.

Credit distribution and limitations for master's degrees. Limitations are placed on the use of credits in 501, 502, 503, 504, 505, 508, and 509 courses. In a 45-credit program, the limits are as follows: a maximum of 12 credits in 501, 502, and 505 combined; a maximum of 9 credits in 504, 508, and 509 combined; a range of 6 to 9 credits in 503. Courses numbered 60x are included in these limitations.

Figure 2: Sidepad document (left) and a PDF mark activated.

We have built some SAs using SPARCE and the mark implementations. The window on the left in Figure 2 shows some SI created using an SA called *Sidepad* [15]. The boxes labeled Level restriction, Credit distribution, and Courses are *items*. The boxes labeled University and Physics are *groups*, which are collections of items and other groups. An item has a name, a comment text, and a reference to base information in the form of a mark (marks are not visible in the figure). A user may double click an item to *activate*, or *see in context*, the mark associated with the item. The window on the right in Figure 2 shows the result of activating the PDF mark associated with the item Credit distribution inside the group University. As a result of activation, the base document is opened in Acrobat, and the referenced sub-document is highlighted. The other three items in the Sidepad document use marks to portions of an HTML document, but those marks too can be activated the same way as the PDF mark. In fact, the Sidepad application, and generally any SA, is agnostic towards the granularity, type, and location of the referenced information; SPARCE manages those details.

With a mark associated with a Sidepad item, a user can also see *context information* such as the text excerpt and page number for the mark from within Sidepad (that is, without activating the mark). Figure 3 shows partial context information displayed inside a *Context Browser* for the PDF mark activated in Figure 2. The text excerpt is currently displayed in the browser. The Context Browser can show several other kinds of context information such as the text excerpt formatted exactly as it is in the base document and the page number.



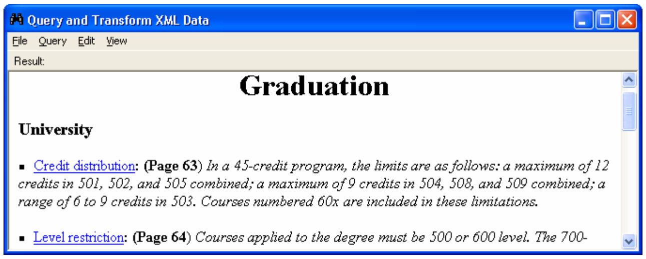
```
In a 45-credit program, the limits
are as follows: a maximum of 12
credits in 501, 502, and 505
combined; a maximum of 9 credits in
504, 508, and 509 combined; a range
of 6 to 9 credits in 503. Courses
numbered 60x are included in these
limitations.
```

Figure 3: Partial context information for a PDF mark.

We have also developed a facility to query and transform the combined superimposed and context information [16]. For example, Figure 4 shows a transformation of the Sidepad document in Figure 2 to an HTML document (using XSLT). This transformation provides an outline view of the Sidepad groups and items. For each item, it includes the name of the item and the text excerpt obtained from the context of the associated mark (shown in italics). The page number where the sub-document resides is shown in bold immediately after the item name. The URL attached to the text containing the item name (denoted by an underline) can be used to activate a mark.

The example in this section demonstrates that SI enables references to heterogeneous information of varying granularity, type, and location, without assuming ownership of referenced information. It shows how marks can help *re-find* information, and demonstrates the ability to see context information without clicking through to referenced information.

The example does *not* demonstrate how SI can help in *finding* information. Also, it assumes that marks are already created for specific information to be re-found. In the rest of this paper, we will discuss how SI can help in finding information, and ways to re-find information without explicitly creating marks to particular sub-documents. Specifically, we focus on the utility of context information in finding and re-finding information.



Graduation

University

- Credit distribution: **(Page 63)** *In a 45-credit program, the limits are as follows: a maximum of 12 credits in 501, 502, and 505 combined; a maximum of 9 credits in 504, 508, and 509 combined; a range of 6 to 9 credits in 503. Courses numbered 60x are included in these limitations.*
- Level restriction: **(Page 64)** *Courses applied to the degree must be 500 or 600 level. The 700-*

Figure 4: Sidepad document transformed to HTML.

3. FINDING INFORMATION

Searching (for example, web search) is a common means to find information, but in both document and sub-document search scenarios, users are often forced to visit (click through) some of the documents and sub-documents to determine their relevance. To help reduce click-through operations, many document search functions these days return a snippet from documents highlighting the occurrence of a search phrase for each document. The search function in Adobe Acrobat does the same for sub-documents. See Figure 1.

The use of marks (or other referencing constructs) and context information can help reduce the number of click-through operations required to determine the relevance of documents or sub-documents. Search functions that return documents based on a search phrase can return a set of marks to the sub-documents that contain (or are otherwise related to) the search phrase. Users can then examine context information for marks without activating the marks. Marks also alleviate the users' need to find the search phrase again within a document because they can use a mark to directly navigate to a sub-document. Also, they can save relevant marks so that they can re-find the same information more easily. Similarly, search functions that return sub-documents also can return marks. Again, users can examine the context of marks to determine the relevance of sub-documents. This approach unifies finding and re-finding activities and can lead to integrated document and sub-document search functions.

4. RE-FINDING INFORMATION

In this section, we outline some of the ways marks and context information assist in re-finding information. We first consider explicitly created marks, and then consider implicitly created marks.

Manually locating a sub-document, especially in a large document, may require considerable user effort. For example, the PDF document in Figure 1 is 368 pages long and the search phrase 'credit distribution' occurs four times in that document. On the other hand, activating a mark takes the user *directly* to the relevant sub-document.

Context information helps users recognize information they have seen before, because they often remember the relative position of a sub-document within a document, or some formatting aspect of the sub-document. Context information can be especially helpful when working with a large collection of marks. (One of the authors of this paper has a collection of over 100,000 marks.) In the example scenario, the user can examine context information to locate candidate marks. She may use either the Context Browser or the query facility to list candidate marks. For example, she can alter the transformation used in Figure 4 to display details for only sub-documents containing the phrase 'credit distribution'. She then can recognize the relevant sub-document using the excerpt, page number, surrounding text, or a formatting aspect such as color.

The means of re-finding we have discussed thus far involve marks already created to specific sub-documents. However, users might wish to re-find information they may have encountered in the process of finding other information. One approach to re-finding is to use *implicit* marks, which are marks expressed intensionally, or marks deduced based on some set of user actions (or inactions). For example, in the scenario described in Section 1, it is possible to mark the search phrase 'credit distribution'. Such a mark is different from the kind of marks we have discussed earlier in this paper because a mark to a search phrase is not a mark to a sub-document; it refers to the collection of sub-documents that result when a search is performed with that phrase. Implicit marks can be generated as "saved searches" as in this example, and by using click-through data from past search results and by monitoring scrolling actions.

5. RELATED WORK

From a human information processing perspective, the use of SI to support re-finding relates to relevant principles involved in recalling information. For example, Wiseman and Tulving [19] originally formulated the encoding specificity principle, which maintains that information items are encoded in our long-term memory (LTM) with respect to their context, and that retrieval is a function of similarity to that encoding context. Thus, retrieval cues, which we use to retrieve information from our LTM, serve as context for information we are trying to recall. In the case of re-finding information, context could refer to many things, including surrounding text, containing document, section heading, and font characteristics.

Studies in PIM have shown that several factors impact re-finding. Capra [2] notes that assessing the future value of information, or post-value recall, can help in information re-finding. Marks, both explicit and implicit, allow users to save information for later, in the original context. Teevan performed a study to explore factors that made search result lists memorable [18]. Her results indicated that factors affecting memorability of results included their rank in the list, whether they were clicked, and other factors like number of times visited. This supports our discussion on implicit creation of marks using click-through data, and their use in re-finding.

Other studies show the importance of context in re-finding. Jones and others [11] found that participants did not use web browser tools, like the bookmarking tool and the history list, to manage information for re-use. Instead they preferred to use methods like emailing web addresses along with comments to themselves and to others. One reason they gave for this was that a self-addressed email can provide an important reminding function together with a context of relevance. Barreau and Nardi [1], in their work on file organization, found that users preferred location-based finding because of its crucial reminding function. In their case, location referred to location of files and not necessarily to bits of information. However, the same idea could be extended so that location of information in its original context serves as a reminder.

Many software systems have been built to improve access to personal information. Hill and Holan describe a system that records history of use of digital objects and later makes use of this history in working with these objects. This use seems to be more in the social context rather than in a personal context [9]. They also enable working with information in parts of documents, with focus on text-based documents. The Mozilla Firefox ScrapBook extension [14] allows users to select bits of information from web pages, save them, and perform searches over this saved information. Although it keeps a reference to the containing web page, it does not retain the context of the selected information within the web page. Gibeo.net network [7] and YAWAS [4] are two other web tools that allow users to select information in web pages, annotate them, view the annotations later, and search over them. These tools limit annotations and re-access to HTML content. Systems like Haystack [12], MyLifeBits [6], Stuff I've Seen [5], and the Google Desktop [8] provide integrated access to more than a single type of information. However, they are limited in their support for working with information at sub-document granularity. The use of SI can enhance current approaches by facilitating reference and access to information at varying levels of granularity, without losing original context.

6. SUMMARY

We have provided an overview of SI and described possible ways of using SI in finding and re-finding personal information. We have illustrated how SI provides context information that may be exploited when finding and re-finding information. We have supported our ideas by tying in relevant work in the PIM field.

Our research to date on SI has provided encouraging evidence that SI may be useful in PIM, but there is scope for improvement. Much of our work to date on SI requires some amount of user involvement and effort (in creating marks). Past discussions about re-finding [10] suggest that re-finding should not require too much effort on the part of users. Although automatic or semi-automatic mark creation (such as creating implicit marks) may be helpful in reducing user effort, deciding which information to automatically mark can be challenging. We also need to evaluate the effectiveness of using SI in finding and re-finding.

7. ACKNOWLEDGMENTS

Sudarshan Murthy's work is supported in part by US NSF grants IIS 0086002 and 0511050. Uma Murthy's work is supported in part by US NSF grant DUE-0435496.

REFERENCES

1. Barreau, D., Nardi, B.A. Finding and Reminding: File Organization from the Desktop. In: SIGCHI Bull. 1995. 27(3) 39-43. See: http://doi.acm.org/10.1145/221296.221307.
2. Capra, R.G. An Investigation of Finding and Refinding Information on the Web. 2005. See: http://scholar.lib.vt.edu/theses/available/etd-03022006-154809/unrestricted/rcapra-dissertation.pdf.
3. Delcambre, L., Maier, D., Bowers, S., Weaver, M., Deng, L., Gorman, P., Ash, J., Lavelle, M., Lyman, J. Bundles in Captivity: An Application of Superimposed Information. In: Proc. of ICDE 2001. 2001.
4. Denoue, L. YAWAS (Yet Another Web Annotation System). 1999-2005. See: http://www.fxpal.com/people/denoue/yawas/.
5. Dumais, S., et al. Stuff I've seen: a system for personal information retrieval and re-use. In: Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003. Toronto, Canada. 72-79.
6. Gemmell, J., Bell, G., .Lueder, R. MyLifeBits: a personal database for everything. In: Communications of the ACM. 2006. 49(1); 88-95.
7. Gibeo. Gibeo Network - Community Web Annotation & Content Aware Tools. 2004. See: http://gibeo.net/.
8. Google. Google Desktop - About. 2006. See: http://desktop.google.com/about.html.
9. Hill, W.C., Hollan, J.D. History-Enriched Digital Objects: Prototypes and Policy Issues. In: The Information Society. 1994. 10(2)
10. Jones, W., Bruce, H. A Report on the NSF-Sponsored Workshop on Personal Information Management, Seattle, WA, 2005. 2005.
11. Jones, W., Bruce, H., Dumais, S. Keeping found things found on the web. In: Proc. of the Tenth International Conference on Information and Knowledge Management. 2001. Atlanta, Georgia, USA. 119 - 126.
12. Karger, D.R., Quan, D. Haystack: a user interface for creating, browsing, and organizing arbitrary semistructured information. In: Proc. of CHI '04 (Extended abstracts on Human factors in computing systems). 2004. Vienna, Austria. ACM Press
13. Maier, D., Delcambre, L. Superimposed Information for the Internet. In: Proc. of WebDB 1999. 1999.
14. Murota Laboratory. ScrapBook - Firefox Extension. 2006. See: http://amb.vis.ne.jp/mozilla/scrapbook/.
15. Murthy, S. Sidepad User Guide. 2005. See: http://sparce.cs.pdx.edu//apps/Sidepad/userguide.
16. Murthy, S., Maier, D., Delcambre, L. Querying Bi-level Information. In: Proc. of 7th International Workshop on the Web and Databases. 2004. Paris, France.
17. Murthy, S., Maier, D., Delcambre, L., Bowers, S. Putting Integrated Information in Context: Superimposing Conceptual Models with SPARCE. In: Proc. of the First Asia-Pacific Conference of Conceptual Modeling. 2004. Dunedin, NZ.
18. Teevan, J. How people recall search result lists. In: Proc. of CHI '06 extended abstracts on Human factors in computing systems. 2006. Montréal, Québec, Canada. 1415-1420.
19. Wiseman, S., Tulving, E. Encoding Specificity: Relation between recall superiority and recognition failure. In: Journal of Experimental Psychology: Human Learning and Memory. 1976. Volume 2. 349-361.