DocuBrowse: Faceted Searching, Browsing, and Recommendations in an Enterprise Context

Andreas Girgensohn[1], *Frank Shipman*[2], *Francine Chen*[1], *Lynn Wilcox*[1]

[1]FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue
Palo Alto, CA 94304, USA

[2]Department of Computer Science &
Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112

{andreasg, chen, wilcox}@fxpal.com, shipman@cs.tamu.edu

ABSTRACT

Browsing and searching for documents in large, online enterprise document repositories are common activities. While internet search produces satisfying results for most user queries, enterprise search has not been as successful because of differences in document types and user requirements. To support users in finding the information they need in their online enterprise repository, we created DocuBrowse, a faceted document browsing and search system. Search results are presented within the user-created document hierarchy, showing only directories and documents matching selected facets and containing text query terms. In addition to file properties such as date and file size, automatically detected document types, or genres, serve as one of the search facets. Highlighting draws the user's attention to the most promising directories and documents while thumbnail images and automatically identified keyphrases help select appropriate documents. DocuBrowse utilizes document similarities, browsing histories, and recommender system techniques to suggest additional promising documents for the current facet and content filters.

Author Keywords

Document retrieval, document management, faceted search, document visualization, document recommendation.

ACM Classification Keywords

H5.2. Information interfaces and presentation: User Interfaces; H3.3. Information storage and retrieval: Information search and retrieval.

General Terms

Algorithms, Design, Human Factors.

INTRODUCTION

Enterprise search has been defined in different ways, including search of an organization's intranet, search of an organization's external web site, and search of any text content in electronic form, such as email and databases [6]. In this paper we focus on search of unstructured information in a corporate document repository. In this type of enterprise search, an employee typically knows or remembers some attributes of the target results [12]. Locating documents in an enterprise often involves finding specific documents that either the user created or the user knows or expects that they were created. In these activities, the user's knowledge of the organization, its history, and its policies and practices can be valuable in locating the desired documents. This is in contrast to Internet users who are searching among a set of unfamiliar documents. Additionally, Internet users are often searching for a fact, such as the phone number for the local restaurant, or a general discussion of a topic, such as what is happening with a particular politician.

In this paper, we present a method for providing search and navigation options to the user through the use of metadata and the document collection file structure. Enterprise document collections are often organized in hierarchies similar to directory trees in file systems. These hierarchies are representations of the policies and practices of the organization, often partly mapping to the structure, roles, and activities in the organization. While newer interfaces allow for access to documents within multiple categories via tags or other mechanisms, the user experience is still relatively similar: users view the set of categories or directories and navigate through the options displayed to locate documents of interest.

To improve system support for navigation and selection in enterprise document collections, we combine data-oriented document analysis with novel interface design. A facet-based interface designed to run in a web browser provides a rich user experience enabling a combination of search and navigation-based location strategies. Because the document type, e.g., spreadsheet, is a useful facet for search, we automatically identify the genre based on features of the page images. Document analysis is also used to determine representative keyphrases of documents to provide a quick overview of each document. These capabilities are part of the new DocuBrowse environment.

In the next section, we elaborate on our vision for enterprise search. The following sections present an overview of the mixed-initiative interface for browsing and searching document collections and discuss recommendation approaches appropriate to the enterprise context. This is followed by a description of the document analysis component with an

IUI'10, February 7-10, 2010, Hong Kong, China.

emphasis on genre identification. We conclude with a vision of how such technologies will change the way people and businesses will store and retrieve documents.

ENTERPRISE SEARCH

Enterprise search is often aided by the user's memory of some properties of the document or the circumstances under which it was presented. We describe two typical examples of this. An employee needs to find information from a past project review presentation. He does not know where the presentation material is located, who gave the presentation, or even which organization in the company created the material. He does recall information in table format in a slide presentation last spring. Another employee is working on solving a problem in product design and remembers a similar problem that was solved in another product several years ago. She would like to find information on the solution, including how it was solved and who solved it. Since the product manager has since left the company, she needs to search the document repository for the information. She knows the product name, the rough time frame, and the nature of the problem.

Unlike in enterprise searches, in typical internet searches the user does not have any prior knowledge of the documents but looks for facts such as a restaurant's phone number or background about a politician. For these types of requests, many different documents meet their needs. In addition to the restaurant's web site, the restaurant phone number could be found on a restaurant review site or a Yellow Pages site. Similarly, articles about the politician can be found on many different news papers and blogs. Another difference is that a Web search result that gets the user close to the result often includes links for browsing to the desired content. In contrast, documents in an enterprise context do not include links with which to browse between documents and that can be used to compute relevance. As the above differences imply, the content-based search techniques used on the Web, while helpful, do not fully address the problem of enterprise search and document access. Thus, other techniques to enhance web-type text queries are needed.

Enterprise search also has to deal with scanned paper documents. While most enterprise documents start in digital form, they often have a period of their life-cycle as paper documents. They need to be signed, annotated, or handed from one person to another. As a result, a document that starts in an easily processable electronic document format (e.g., Microsoft Word) often becomes a scanned document later. The content of the scanned document may be mostly reconstructed through OCR but the result is that all types of documents, whether they start out in a word processor, a spreadsheet, a database, or presentation software, end up as the same document format.

Recommender systems are useful in many contexts and have become common tools for people finding movies, books, etc. Thus, it is natural to apply recommender techniques to the enterprise context. Unfortunately, most recommender techniques are not directly applicable due to assumptions about information availability and access homogeneity. Typically, recommender techniques rely on individuals having access to the same set of resources. In a corporate context, each employee is likely to have access to a different subset of resources. This may undermine some of the statistical analysis techniques commonly used. These systems also rely on users being willing to share evaluations of resources and interests. Such information is unlikely to be available in the enterprise context. The enterprise social setting makes it inappropriate for employees to rate each other's (or their boss's) documents. Likewise, issues of privacy and compartmentalization makes it unlikely for information that can be used to determine who is working on what to be centralized. However, we can use certain types of recommendations such as documents matching the query that other users with similar queries viewed or documents that were recently viewed by other users in the same organization.

DOCUBROWSE

Our DocuBrowse system supports faceted searching, browsing, and recommendations in an enterprise context. It combines well-known techniques for supporting document access, including browsing the structure of the document collection, searching content, filtering based on metadata, and presenting recommendations based on past user or group activity (see Figure 1). We have deployed the system internally to provide access to 9,000 office documents and 50,000 images created over the past ten years.

Browsing the Document Collection

Enterprise workers typically organize their document collections into sub-collections that group documents based on characteristics of their content, generation, or use. In file system-based document stores, the collection structure is the directory hierarchy. For documents with metadata, documents may be placed based on an ontology of the metadata concepts (e.g., MeSH Terms for the National Library of Medicine).

Acknowledging the centrality of browsing through collections, we created DocuBrowse as a web-based interface that makes browsing such structures easy and intuitive. As such, the majority of DocuBrowse's display is used to present the contents of the user's current location in the document collection. Subcollections (e.g. directories) are presented above the individual documents. DocuBrowse employs modern web technologies to quickly respond to user requests in an asynchronous fashion, for example, by retrieving additional information about a document to be displayed in a tool tip.

Unlike traditional file systems, DocuBrowse allows users to put a document into more than one "directory" and, unlike Windows short cuts, each document entry provides direct access to the original document. In the DocuBrowse prototype, documents are identified by a content hash. This enables automatic duplicate detection so that only one copy of the document has to be stored. As a change to a document changes its identifier, all entries for that document have to be updated. A history mechanism can point to previous versions of a document. To explicitly point to an older version of a document, an entry has to be marked to prevent updates for new document versions.

The document hierarchy is stored in a database. To speed up queries that need to check all documents in a directory sub-

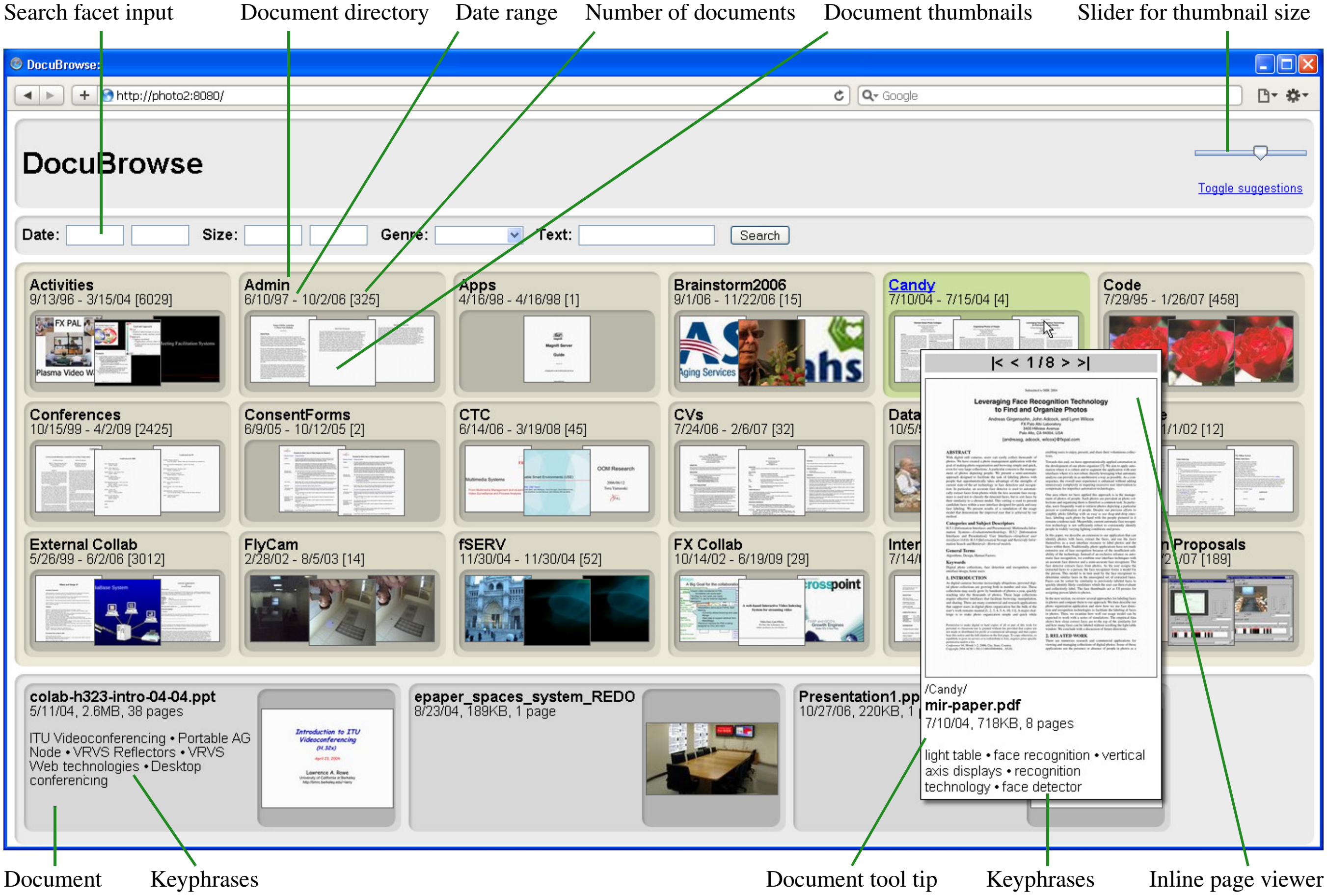


Figure 1. Top-level of the DocuBrowse document hierarchy.

tree, e.g., for selecting thumbnails for directories, we also store the containment closure by directly storing all parent directories for each document and directory. That enables us to retrieve all documents or directories in a directory tree with a single query. While the initial document hierarchy is modeled after the file system hierarchy provided by the user, multiple hierarchies can be supported. Each hierarchy is stored in database tables so that no changes to the flat document storage area are required when a hierarchy is added or changed. Other hierarchies may be based on document properties or on external semantic hierarchies such as the Library of Congress classification for books.

Visualizing Documents

Documents are visualized by a box including the document name, metadata, and automatically selected keyphrases on the left and a thumbnail of the first page of the document on the right (see Figure 2). Clicking anywhere in the document opens the document in the document viewer. The document thumbnails and the information text are offered in different sizes that can be changed quickly by using a slider (see Figure 1). This slider employs dynamic web technologies to zoom in or out without having to reload the whole page. Thumbnails are created in many different sizes when documents are added to the collection so that thumbnails of a requested size can be shown quickly. Figure 4 shows a page with larger thumbnails.

Subcollections, or more simply collections, are visualized by three thumbnails of selected documents in that collection (see Figure 2). Those thumbnails are cropped to squares to better fill the directory box. If fewer than three documents

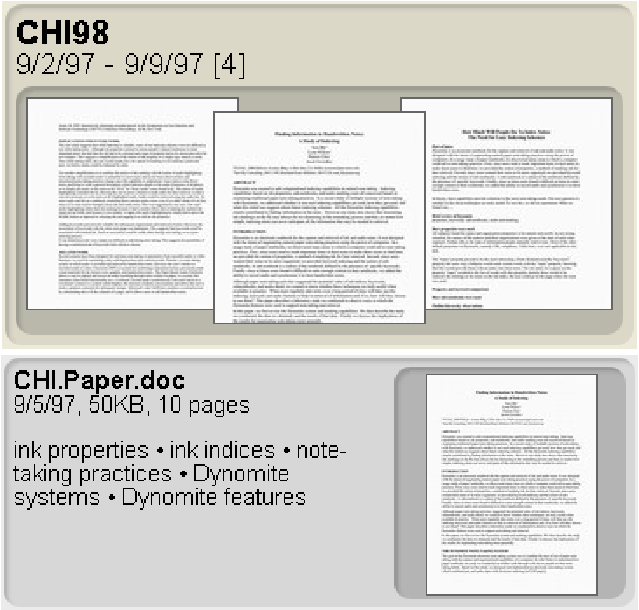
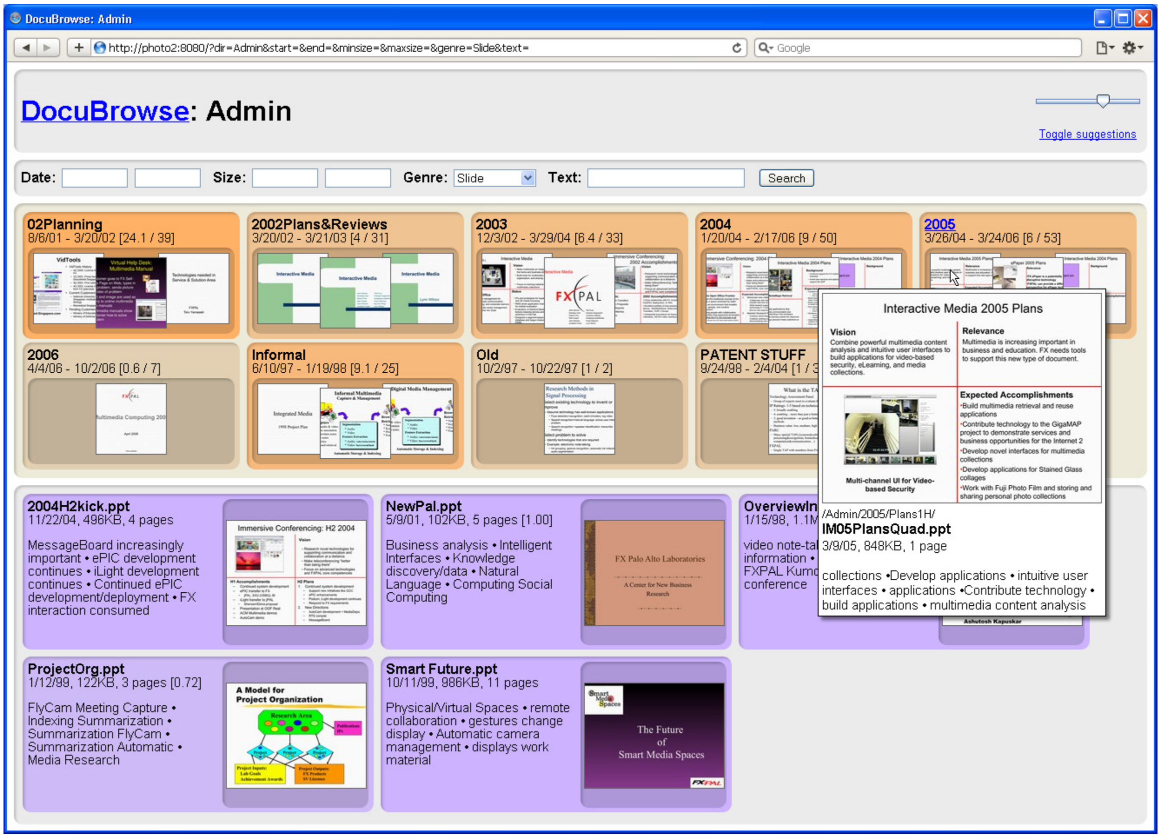

CHI98
9/2/97 - 9/9/97 [4]
CHI.Paper.doc
9/5/97, 50KB, 10 pages
ink properties • ink indices • note-taking practices • Dynomite systems • Dynomite features

Figure 2. Directory and document representations.



DocuBrowse: Admin

http://photo2:8080/?dir=Admin&start=&end=&minsize=&maxsize=&genre=Slide&text=

DocuBrowse: Admin

Toggle suggestions

Date: Size: Genre: Slide Text: Search

02Planning
8/6/01 - 3/20/02 [24.1 / 39]

2002Plans&Reviews
3/20/02 - 3/21/03 [4 / 31]

2003
12/3/02 - 3/29/04 [6.4 / 33]

2004
1/20/04 - 2/17/06 [9 / 50]

2005
3/26/04 - 3/24/06 [6 / 53]

2006
4/4/06 - 10/2/06 [0.6 / 7]

Informal
6/10/97 - 1/19/98 [9.1 / 25]

Old
10/2/97 - 10/22/97 [1 / 2]

PATENT STUFF
9/24/98 - 2/4/04 [1 / 3

2004H2kick.ppt
11/22/04, 496KB, 4 pages

MessageBoard increasingly important • ePIC development continues • iLight development continues • Continued ePIC development/deployment • FX interaction consumed

NewPal.ppt
5/9/01, 102KB, 5 pages [1.00]

Business analysis • Intelligent Interfaces • Knowledge discovery/data • Natural Language • Computing Social Computing

OverviewIn
1/15/98, 1.1M

video note-ta information • FXPAL Kum conference

ProjectOrg.ppt
1/12/99, 122KB, 3 pages [0.72]

FlyCam Meeting Capture • Indexing Summarization • Summarization FlyCam • Summarization Automatic • Media Research

Smart Future.ppt
10/11/99, 986KB, 11 pages

Physical/Virtual Spaces • remote collaboration • gestures change display • Automatic camera management • displays work material

Interactive Media 2005 Plans

Vision
Combine powerful multimedia content analysis and intuitive user interfaces to build applications for video-based security, eLearning, and media collections.

Relevance
Multimedia is increasing important in business and education. FX needs tools to support this new type of document.

Multi-channel UI for Video-based Security

Expected Accomplishments
•Build multimedia retrieval and reuse applications
•Contribute technology to the GigaMAP project to demonstrate services and business opportunities for the Internet 2
•Develop novel interfaces for multimedia collections
•Develop applications for Stained Glass collages
•Work with Fuji Photo Film and storing and sharing personal photo collections

/Admin/2005/Plans1H/
IM05PlansQuad.ppt
3/9/05, 848KB, 1 page

collections •Develop applications • intuitive user interfaces • applications •Contribute technology • build applications • multimedia content analysis

VidTools
Virtual Help Desk: Multimedia Manual
Technologies needed in Service & Solution Area
Interactive Media
FX PAL
Immersive Conferencing: 2002 Accomplishments
Interactive Media 2004 Plans
Interactive Media 2005 Plans
ePaper 2005 Plans
Multimedia Computing 2006
Integrated Media
Informal Multimedia Capture & Management
Digital Media Management
Research Methods in Signal Processing
Immersive Conferencing: H2 2004
FX Palo Alto Laboratories
A Center for New Business Research
A Model for Project Organization
Research Area
The Future of Smart Media Spaces
Ashutosh Kapuskar

Figure 3. Subcollection restricted to documents matching the "slide" genre.

exist in that collection, then only that number of thumbnails is presented. Thumbnails are selected such that a good sample of the documents in the collection is provided. If a keyword or facet search is active, directory thumbnails are chosen among the documents that best match the search query. In addition to the thumbnails, the collection's name and statistics about its contents are shown. The statistics include the number of documents and the date range. Clicking on the collection visualization navigates to that subcollection, replacing the current list of collections and documents.

When the mouse lingers over a document or a thumbnail for a document in a subcollection, an interactive tool tip appears that provides easy access to more detailed information about that document. The tool tip also includes a small document viewer that allows the user to flip through images of the document pages. This lets the user quickly verify if this is the desired document before opening the full-size document viewer.

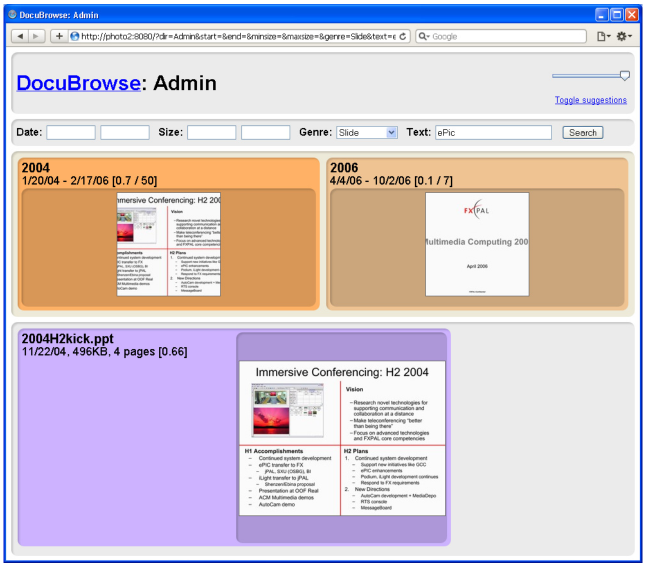
Search- and Filter-Based Navigation Support

Unlike traditional search systems that display a list of matching documents, DocuBrowse uses a combination of filtering, color-coding, and browsing to present search results. This approach keeps matching documents in context and makes it easy to narrow or widen a search for a subcollection. Because the document organization tends to reflect the structure and practices of organizations, we maintain that structure in the visualization of results.

To find documents with certain characteristics or contents, DocuBrowse provides filters for different facets of the documents. An important facet is the type, or genre, of a document. Our automatic genre detector is trained to identify technical papers, slides, tables, and photos that correspond roughly to the file formats MS Word, PowerPoint, Excel, and JPEG. By selecting a genre, only documents matching the genre and directories containing those documents are shown. For fuzzy genre matches, color coding is used to indicate the strength of the match.

In addition to genres, DocuBrowse can filter the results based on document size and date. More options can be provided for document collections where additional metadata facets are available. When the user specifies values for a facet, only those documents that fit those values are shown in the browser. DocuBrowse also supports full-text search, so documents that partially or completely match the terms in the query are displayed in the browser while non-matching documents are filtered out. As with fuzzy genre matches, color coding indicates the quality of the match.

When a search or filter is active, the visualization of subcollection is colored to show where large numbers of matching documents are located (see Figure 3). The score for a directory tree combines the score of the best-matching document with the total number of matching documents and the density of the match scores. The match score for a directory is given by:



DocuBrowse: Admin

http://photo2:8080/?dir=Admin&start=&end=&minsize=&maxsize=&genre=Slide&text=e

DocuBrowse: Admin

Toggle suggestions

Date: Size: Genre: Slide Text: ePic Search

2004
1/20/04 - 2/17/06 [0.7 / 50]

2006
4/4/06 - 10/2/06 [0.1 / 7]

2004H2kick.ppt
11/22/04, 496KB, 4 pages [0.66]

Immersive Conferencing: H2 2004

Figure 4. Slides containing the text "ePIC."

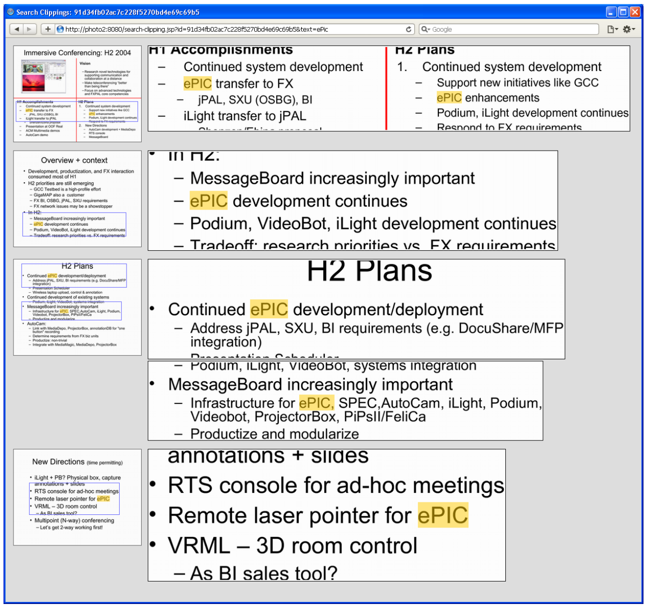
$$s = b \cdot \sqrt{\frac{d^2 + c^2}{2}}$$

where b is the best match score among documents in the directory tree, d is the normalized density, and c is the normalized count. The density is the average match score, including documents with a match score of zero. The count is the number of documents with a non-zero match score. Both d and c are normalized relative to the greatest value from the subdirectories being compared. For combining d and c, we chose the quadratic mean because it comes close to picking the maximum of the two values without completely ignoring the other value. Document thumbnails for each directory are chosen from those documents that best match the query while balancing the selection from different branches of the directory tree. Documents with multiple paths to them are only included once.

Multiple document facets can be combined to further restrict the matching documents. For example, after having located slides in a document collection, the user can further restrict the matching documents to those that also contain the specified text string (see Figure 4). Note that each directory in Figure 4 is only visualized by a single document thumbnail even though multiple documents are contained in those directories. This is due to the fact that only a single document matches the query in each of those directories. For combining the results of fuzzy facet searches, we multiply the individual document scores. That produces results consistent with conjunctions in boolean contexts.

Document Viewer

The DocuBrowse viewer provides access to the document pages without requiring other software such as Flash or Adobe Acrobat. It offers two fairly traditional views. One view provides thumbnails of all pages in the documents. Just like in the collection view, a slider allows the user to quickly change the size of the thumbnails. While such a view is available in applications such as Microsoft PowerPoint, it is less common in document viewers. The reading view includes thumbnails on the left with a large view of a single page on the right. It is quite similar to a view provided by Adobe Acrobat.



Search Clippings: 91d34fb02ac7c228f5270bd4e69c69b5

http://photo2:8080/search-clipping.jsp?id=91d34fb02ac7c228f5270bd4e69c69b5&text=ePic

H1 Accomplishments
– Continued system development
– ePIC transfer to FX
– jPAL, SXU (OSBG), BI
– iLight transfer to jPAL

H2 Plans
1. Continued system development
– Support new initiatives like GCC
– ePIC enhancements
– Podium, iLight development continues
– Respond to FX requirements

In H2:
– MessageBoard increasingly important
– ePIC development continues
– Podium, VideoBot, iLight development continues
– Tradeoff: research priorities vs. FX requirements

H2 Plans
- Continued ePIC development/deployment
  – Address jPAL, SXU, BI requirements (e.g. DocuShare/MFP integration)

– Podium, iLight, VideoBot, systems integration
- MessageBoard increasingly important
  – Infrastructure for ePIC, SPEC,AutoCam, iLight, Podium, Videobot, ProjectorBox, PiPsII/FeliCa
  – Productize and modularize

annotations + slides
- RTS console for ad-hoc meetings
- Remote laser pointer for ePIC
- VRML – 3D room control
  – As BI sales tool?

Figure 5. Clippings of pages containing search text.

The snippet view displays more details for full text search results. It shows the thumbnails of pages with matching terms and a larger view of the snippets of text in which those terms appeared (see Figure 5). This view provides a quick view of text matches that are too small to see in a thumbnail view. By showing the outlines of snippets in the page thumbnails, context for the snippets is provided.

Search Scenario

To illustrate our approach to search, we describe a use scenario. The Director of Research at a research laboratory wants to access details about a presentation system that was created a few years ago. When looking at the top-level view presented by DocuBrowse (Figure 1), she notices the "Admin" directory and realizes that project reviews stored in that directory would be a good source for the desired information. After navigating to that directory, she restricts the view to documents in the "slide" genre (Figure 3). Among the keyphrases displayed for one of the documents is the word "ePIC" that sounds familiar. To make sure that this is indeed the desired system, she performs a text search with that term. This reduces the view to two subcollections and one document (Figure 4). Clicking on the document displays search clippings that provide sufficient information to identify the system (Figure 5).

ENTERPRISE-ORIENTED RECOMMENDATIONS

Designing mechanisms to generate suggestions requires an understanding of the structure of activity within an enterprise. We look at different properties of corporate organizations. Once appropriate recommender groups have been identified, we base our recommendations on recency, type of access, and document similarity. We present recommendations within DocuBrowse and indicate the basis for each recommendation.

Defining Recommender Groups in a Corporate Setting

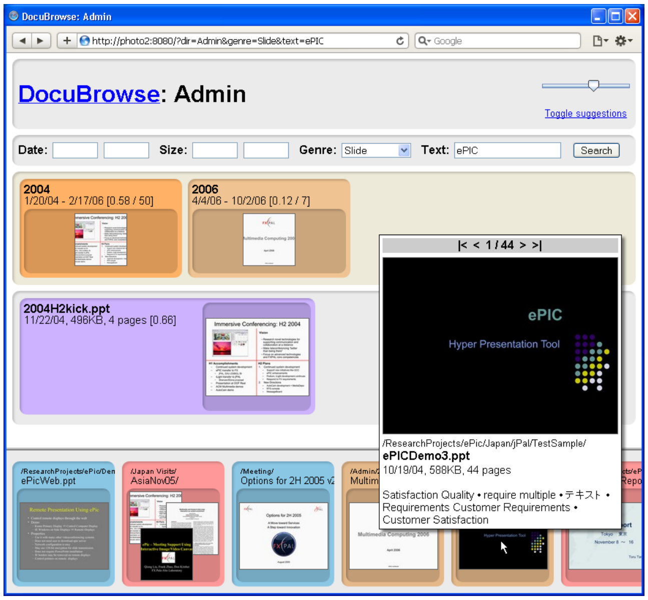
As described by Simon, the hierarchic structures of organizations are meant to limit the need for information flow between parts of the organization [15]. As a result, only very general documents such as phone lists, policy descriptions, and guidelines are likely to be widely available across an organization. This raises the problem of identifying activity in the organization's information access that is predictive of future needs of an individual.

Instead of using interaction history of the whole user community to recognize people with similar information needs, as is generally true in recommender algorithms, we propose to use subgroups of the organization chosen based on an understanding of information access in organizations. Two attributes of individuals are being used to identify subgroups:

- *Organizational structure.* As the basis of Simon's argument about the effects of bounded rationality on organizational structure, the information needs of people in the same part of the organization are likely to be indicative of the needs of others in that part of the organization. The first subgroup considered are those individuals that are part of the same organizational component. Determination of the organizational levels used for this grouping requires knowledge of the organization.
- *Job classification.* Simon notes that the organizational structure is not the only hierarchic decomposition of an enterprise. A classification indicative of the type of activity one is involved in is one's job title and different types of activities (e.g. accounting, purchasing, administration, etc.). Thus, our second subgroup used to generate suggestions is the set of people with the same or similar job title. Again, some knowledge of the organization is required to determine which job titles (e.g. assistant professor, associate professor, professor) should be combined into a single group.

To generate suggestions based on each of these groups, we aggregate individual access histories into relevance scores as described below. Thus, all of the individuals in a specified organizational layer are included in organizational structure recommendations. If a complete organizational structure is available, the individual assessments can be weighted by distance from the individual accessing the document store. Otherwise, all individuals within that structure are equally weighted. Similarly, all individuals within the set of job titles defined as equivalent for this purpose are included for generating job classification recommendations.

A final change with regard to traditional recommender algorithms is that suggestions can be for directories as well as individual documents. Directories are important in organizations as the location where documents in a particular sequence are kept. Thus, while past interactions with the January, February, and March accounting files for a office would not point to the April accounting file in a traditional recommender approach, they will point to the directory that includes the April file.


DocuBrowse: Admin

http://photo2:8080/?dir=Admin&genre=Slide&text=ePIC

DocuBrowse: Admin

Date: Size: Genre: Slide Text: ePIC

2004
1/20/04 - 2/17/06 [0.58 / 50]

2006
4/4/06 - 10/2/06 [0.12 / 7]

2004H2kick.ppt
11/22/04, 496KB, 4 pages [0.66]

|< < 1 / 44 > >|

/ResearchProjects/ePic/Japan/jPal/TestSample/
ePICDemo3.ppt
10/19/04, 588KB, 44 pages

Satisfaction Quality • require multiple • テキスト • Requirements Customer Requirements • Customer Satisfaction

/ResearchProjects/ePic/Den
ePicWeb.ppt

/Japan Visits/
AsiaNov05/

/Meeting/
Options for 2H 2005 v2

/Admin/
Multim

Google
Toggle suggestions
Search
ePIC
Hyper Presentation Tool
Immersive Conferencing: H2 2004
Multimedia Computing 2006
Remote Presentation Using ePic
Options for 2H 2005
FXPAL

Figure 6. Suggestions for relevant documents.

Computing Document Value

Our recommendations are provided based on recency, type of access, and document similarity. There are three methods to generate suggestions. The first is based on the individual's past interactions with documents. The second and third are based on access and viewing by the subgroups of the organization just described. To compute the likely interest value of a document for a particular individual, we currently use a simple model to incorporate the type and history of access by that individual. DocuBrowse distinguishes between documents that have been viewed in the interactive tool tip and those that were opened in the document viewer. We consider the former to be a quick glance and the latter a more detailed exploration. To differentiate between those types of access, we record the most recent view event for each and then apply an exponential decay to an initial score. In our current implementation, both types start with the same score. Tool tip views have a half life of 30 minutes and detailed views have a half life of 60 minutes. This framework can be expanded to include additional forms of access as they become available in DocuBrowse. Also, instead of only considering the most recent view, more of the interaction history could be considered. For example, the decaying scores for all views could be added.

Computing the interest value of a directory is based on the interest of the files within that directory. When a significant fractions of documents in a directory would be suggested, we instead suggest the whole directory. This approach is applied recursively such that the parent directory is recommended if most subdirectories and documents in it are deemed to be relevant.

Presenting Suggestions in DocuBrowse

DocuBrowse presents suggestions in a floating pane below the main browsing interface (see Figure 6). If a directory is suggested instead of an individual document, the directory is visualized in a fashion similar to the directory listing. The thumbnails of the three most relevant documents are shown with the most relevant one in the front (see second suggestion in Figure 6).

Because there are multiple methods used to generate suggestions, DocuBrowse exposes the form of reasoning used to generate each suggestion through color coding. We currently indicate the following types of suggestions:

- Suggestions based on personal interaction history.
- Suggestions based on the interaction history of members of one's organizational branch.
- Suggestions based on the interaction history of employees with similar job titles.
- Suggestions based on multiple lines of reasoning.

Users of the system do not need to know the specifics of the reasoning approaches. It is natural that some forms of reasoning are more valuable for some jobs than others. Experiences examining suggestions with different color codings will lead users to learn which classes of suggestions work best for them.

KEYPHRASE SELECTION

Keyphrases that give a sense of the content of a document are used in the document summaries presented in DocuBrowse. A small number of keyphrases that can be compactly presented are needed. We decided that a larger number of shorter keyphrases would be more informative, and so five keyphrases that are up to three words long are selected for each document.

Our method identifies sequences of words between stop words and other textual cues, such as punctuation, including PowerPoint bullets, and changes in font style and size, as candidate keyphrases. For each document, the candidate keyphrases are scored and the best *N* keyphrases selected, where *N* is prespecified and may be dependent on the amount of screen space available to the application.

To select the best keyphrases, a weighted combination of features is used. The features are text based and include: (1) number of times a term occurs in the document, (2) number of documents in which a term occurs at least once in an English corpus, (3) number of tokens in the keyphrase, (4) location of first mention of the term in the document, measured as paragraph number.

The weighted combination of features is given by:

$$Score(k_j) = \sum_i \lambda_i f_i(k_j, d)$$

where λ_i is the weight given a feature and $f_i(k_j,d)$ is the value of feature i for keyphrase candidate k_j in document d. Once each of the keyphrases is scored, they are then ranked against each other and the best keyphrases are selected for each document.

Evaluation of the keyphrase selection has been limited to visual inspection of the keyphrases displayed by DocuBrowse (Figures 1, 2, and 3). As with sentence-based summarization tasks, there are generally many more keyphrases that are suitable keyphrases for conveying the gist of a document than can be displayed, and evaluation is a tricky endeavor. For technical papers and slides, the keyphrases displayed by DocuBrowse have been found to be good.

Although we have described keyphrase selection by document, the approach can be applied to other units of text. Our keyphrase selection system has been used to select keyphrases for each page of a document. In these cases, the keyphrases are selected within a document, and additional features and methods are used to reduce redundancy in the selected keyphrases. On the other hand, our keyphrase selection system could also be used to select keyphrases for larger units, such as a subdirectory, although the usefulness would depend on the coherence of the documents in each directory.

GENRE IDENTIFICATION

Documents are often classified and searched for based on words and topics. However, documents can also be classified by another independent attribute: genre. Example genres in literature include poetry, fiction, and drama. In enterprise search and browsing, a different set of document genres than those used to describe literature are needed.

Documents in the DocuBrowse repository are created using a variety of tools that produce documents in different formats, including PDF, DOC, XLS, JPG, and PPT. In a simplistic approach to genre identification, the different document formats roughly correspond to different genres, and a document filename extension could be a surrogate for document genre. However, a document creation tool is often used to create documents in more than one genre. In addition, some extensions, such as PDF, are associated with many genres, since files created in different formats are often converted to PDF for its portable representation. Thus, using only file extensions would result in a user looking for slides to see PowerPoint files that are not slides. Furthermore, the user will not see slides that are in other formats, such as PDF. Similarly, scanned document pages may be JPEG files, so that their genre is unknown.

In genre identification for DocuBrowse, the documents are automatically categorized into a small number of genres, roughly corresponding to the genres usually associated with document format types: technical paper, slides, table, and photo. Figure 7 shows examples of pages from each of the four genres that our Genre Identification and Estimation system (GenIE) has been trained to identify. Note the variation within a genre and that, while the text might not be large enough to read, the genre of each page is readily apparent.

The GenIE System

In the GenIE system, documents are classified based on features extracted from page images. Those images are either generated by scanning paper documents or by rendering electronic documents. Our approach tries to capture layout features without explicitly performing layout analysis. In particular, each page image is tiled and document-based image features are extracted to characterize each tile. Genre identification is performed per page, and then document genre is estimated based on the genre estimates for the pages in the document.

A Comparison of Software and Hardware Techniques for x86 Virtualization
Infectious disease risks from dead bodies following natural disasters
Implementing an Untrusted Operating System on Trusted Hardware
WHAT THE DOG SAW
Mixed Media Interactions
Human-Computer Interaction Consortium
Founded in 1988
Fosters interaction among universities, industries, and government research laboratories interested in HCI
15 members (8 academic; 6 industry)
Interactive Media
Rich Media Organizer (RMO)
Collaboration with FSCA
Segment-based Multimedia Storage and Retrieval
Architecture request from IML
Hyper-Hitchcock
ADO: Audio Device Organizer

Figure 7. Sample document pages from the genres of technical papers, photos, tables, and slides.

We developed GenIE using a corpus of documents crawled from the web, some of them printed and scanned by us, and some directly converted to JPEG. GenIE was trained to identify the four genres of interest: technical papers, slides, photos, and tables.

We developed a set of features that capture local document characteristics, such as lines of text or text size, within a tile. The tiles must be large enough to extract document characteristics within each tile and at the same time small enough so that the different region types (e.g., heading, figure, body text) remain distinct. Empirically, we have found that dividing each page image into a grid of 5 tiles horizontally by 5 tiles vertically, for a total of 25 tiles, meets our requirements.

The following features are computed for each tile: (1) image density [10], (2) horizontal pixel projection, (3) vertical pixel projection, and (4) color correlogram. Three page-based features are computed: (1) horizontal line lengths, (2) vertical line lengths, and (3) image size.

Each document may be tagged with zero or more genres. To handle tagging with multiple genres, a separate classifier was trained for independently identifying each genre. In developing the genre identification classifiers, a corpus of documents composed of 5098 pages from 1081 documents was labeled with the targeted genres with a total of 3670 labels. Because of the competitive performance of support vector machines (SVM) for many classification tasks, we used an SVM classifier, SVMlight [8]. A separate classifier using a one-against-many model was trained for each genre. We reimplemented Kim and Ross' algorithm [10] as a baseline for comparing performance of our GenIE system and observed that GenIE performed noticeably better (mean precision of 0.59 vs. 0.94 and F1 of 0.51 vs 0.84 for Kim and Ross vs. GenIE, respectively).

To tag the documents in the DocuBrowse corpus, the four GenIE document genre identifiers were used to score each page of all the documents in the corpus. For the DocuBrowse task, the emphasis is on high recall, and our method of combining the page scores for tagging documents by genre bears this in mind. A document genre score, $S_d(g)$, for document d being genre g is computed for each genre as the average of the individual page scores for a document. A page score is the averaged SVM score, clipped to a minimum value of 0.0 and a maximum value of 1.0.

$$S_d(g) = \frac{1}{2P}\sum_{p,\,c} max(min(s(p,\,g,\,c),\,0.0),\,1.0)$$

where $s(p,g,c)$ is the SVM score for page p being classified as genre g by classifier c and P is the total number of pages in a document. We used two classifiers per genre where each classifier was trained on a separate data partition.

In addition to GenIE's good performance in comparison to a baseline implementation, we also found that it provided a major contribution in the context of DocuBrowse. The offered genres are intuitive and filter the document collection in a useful fashion. For example, restricting the document hierarchy to just slides is very helpful for finding project presentations.

RELATED WORK

There has been much work in several areas related to our work presented here. In this section, we highlight how our work draws from and extends this earlier work, and how we tailor it to address the characteristics of enterprise search.

Document Browsers and Faceted Search

Facets have been presented and used in search and browsing systems in a variety of ways, and the set of facets supported varies, depending on the contents of the document repository. A study by Wilson and schraefel [18] found that "a balance of exploratory and keyword searches" was performed using the mSpace faceted browser both during early use of the system and in later use, indicating that both browsing and search should be supported. This is a feature in both our system and the faceted search systems that we will contrast with ours.

The mSpace faceted browser lays out categorical facet values in columns which can be moved to indicate the priority of filtering relationships. Our system also provides for user-ordered filtering by facet, but the ordering is determined by the order that a user specifies facet values of interest.

Hearst [5] provides recommendations for the design and layout of hierarchical, categorical facets in the Flamenco system and also comments on the use of facets in the eBay Express interface. In Flamenco, which was developed on a collection of fine art images, the result set is shown on the right half of the interface, with result items grouped by the most recently selected facet.

Microsoft's FacetLens system [11] lays out facets and their attribute values in rectangular regions, with the result set also presented on the right side of the interface. FacetMap [16], a predecessor to FacetLens, was also developed at Microsoft for personal information stores with rich metadata. It uses facets for organizing dataset items and dynamically allocates screen space based on the distributions of attributes among the search result set. FaThumb [9] was also developed at Microsoft and provides faceted mobile search.

In contrast with these systems, our system relies on the directory hierarchy to provide the grounding context for users, and results are presented by filtering out parts of the directory hierarchy and by highlighting directories to indicate those that are good matches. In addition, the facets in our system are a mix of pre-specified categories and facets with many values, such as a date range or range of file sizes.

These earlier systems were developed for a browsing task with a relatively homogenous collection of item types, while our "enterprise" document collection contains a variety of document types, such as technical papers and slides. The UpLib system [7] is a "personal digital library system" that also contains a variety of document types, such as technical papers and slides, similar to our document corpus. In UpLib, metadata can be stored with each document. Documents are accessed either by viewing all the documents in the repository or by search over the text and metadata, such as "authors" or "keywords" using the Lucene system. The search-based approach is in contrast to the faceted systems that prompt the user to select a facet value and so reduce the cognitive load on the user.

DocuBrowse draws from the work of these earlier faceted search and browsing systems. Each of these earlier systems was developed for collections with rich metadata that is a significantly stronger organizational factor than the file directory hierarchy. For our task context of enterprise collections, the file hierarchy plays a significant role, and we developed DocuBrowse to integrate faceted search and browsing with a file hierarchy. As with the other faceted systems, facet selection narrows the repository items presented. But unlike earlier systems, the presentation of documents is integrated with the file hierarchy.

Enterprise-Oriented Recommendations

DocuBrowse also provides recommendations, unlike the earlier faceted search systems. Although recommendation systems for different types of data have been developed, we have not seen a recommendation system for enterprise collections. DocuBrowse uses enterprise-based subgroups to generate different types of suggestions. Abecker et al. [1] present an information architecture designed to support the active delivery of resources in an enterprise but do not explore the availability of information necessary to support the active delivery process. Plu et al. [13] describe using recommender techniques to suggest contacts in a corporate environment. More related to DocuBrowse, Zhen et al. [20] focus on fact-oriented recommendations, such as experiences with a particular vendor or with a specific product.

Keyphrase Selection

Although some documents contain keyphrases, many do not. The automatic selection of document keyphrases provides for greater coverage in the use of keyphrases when visualizing documents. There are a number of ways to identify keyphrases (e.g., [17]). A straight-forward method is by tagging the part-of-speech (POS) of the text and then identifying POS tag sequences that correspond to a noun phrase [17]. Another method is to identify sequences of words between "stop words," or non-content words [3]. More recently, supervised methods which learn how to combine different features have been successfully used. One such system is the KEA system [19]. These systems require a training set of documents labeled with keyphrases. Since we do not have a labeled corpus and it would have been time-consuming to manually label a reasonable size corpus for training a system, we took an unsupervised approach that produces reasonable keyphrases.

Genre Identification

The data collections used by the earlier systems were rich in metadata, where much of the metadata was manually entered. For enterprise collections, the work practice of employees generating documents usually does not include entering metadata for each document. Additionally, when metadata is entered in an ad hoc manner by many people, it is often inconsistent. By automatically extracting metadata, such as genre, from a document collection, more consistent metadata is available for use in faceted search. Although genre identification systems have been developed, the genres covered were not a good match to our enterprise collection and the performance was not good enough for use as facets. Automatic genre identification based on text and markup features has been proposed for web search improvements [2]. Other systems have been developed for classifying page images into different genres.

Shin and Doermann [14] perform document layout analysis to label and identify the boundaries of different types of document regions (e.g., text, image, graphics) and extract features for use in a decision tree classifier. Their identified "genres" correspond to page types: cover page, reference, table of contents, and form pages.

Although layout analysis can be successfully performed for limited domains, general layout analysis is still not robust. Two other image-based approaches to genre identification that do not require layout analysis have been developed by Gupta and Sarkar [4] and by Kim and Ross [10]. Gupta and Sarkar identified salient feature points and performed classification based on the points' locations and local image characteristics. However, their genre classifier was tested on discriminating between only two types of genres: journal articles and memos.

Kim and Ross [10] developed an image-based genre classifier for the first page of a document. They divide the page into a uniform grid of 62 by 62 tiles, count the number of non-white pixels in each tile to compute black pixel density, and use the tile densities as classifier features. They compared the performance of the count feature using Naïve Bayes, Random Forest, and support vector machine (SVM) classification methods. Their best-performing image-based genre classifier performance was Naïve Bayes, but the performance was relatively poor and is meant to be used in conjunction with a text-based genre classifier.

FUTURE WORK

Our system has been in internal use for several months to provide access to historical research documents of our laboratory. We have anecdotal evidence that our approach for combining searching and browsing helps in locating documents that would be missed in a pure search. For example, we located slides depicting Japanese visitors from a previous visit ten years ago. We do not have a sufficient history of interactions with this set of research documents to fully evaluate our algorithms and interfaces for recommending documents. Furthermore, our laboratory is small and has a flat organization, so we could not test recommendations based on organization. We are currently exploring ways to deploy DocuBrowse in a larger organization and are determining requirements for additional security measures and tools.

CONCLUSIONS

We presented a new approach for searching and browsing enterprise document collections that combines faceted filtering and search with navigation of the document collection structure. These techniques are part of a web-based system for accessing document collections. Unlike traditional document search systems, our system presents search results within the user-created document hierarchy by only showing matching documents and directories and highlighting promising areas. To support scanned-in documents and to better classify electronic documents, we utilize an automatic genre identifier that can determine genres such as papers, slides, tables, and photos. We also automatically determine keyphrases that provide users with a quick overview of the document content. We look at access histories, job roles, and document similarities to recommend additional documents that may be useful to the user. We expect this novel approach to simplify access to enterprise document collections.

REFERENCES

1. A. Abecker, A. Bernardi, K. Hinkelmann, M. Sintek. Enterprise Information Infrastructures for Active, Context-Sensitive Knowledge Delivery, in *Knowledge Management Systems: Theory and Practice*, S. Barnes (ed.), Thomson Learning, pp. 146-160, 2002.
2. E.S. Boese and A.E. Howe. Effects of web document evolution on genre classification. In *Proc. of ACM CIKM 2005*, pp. 632-639, 2005.
3. F. Chen, S. Putz, D. Brotsky. Automatic method of selecting multi-word key phrases from a document. US Patent 5745602.
4. M.D. Gupta and P. Sarkar. A shared parts model for document image recognition. In *Proc. of the Ninth International Conference on Document Analysis and Recognition*, pp. 1163-1172, 2007.
5. M. Hearst. Design Recommendations for Hierarchical Faceted Search Interfaces. In *Proc. of the ACM SIGIR Workshop on Faceted Search*, pp. 26-30, 2006.
6. D. Hawking. Challenges in Enterprise Search. In *Proc. of Australasian Database Conference,* pp. 15-24, 2004.
7. W. Janssen and K. Popat. UpLib: a universal personal digital library system. *Proc. of ACM Symposium on Document Engineering*, pp. 234-242, 2003.
8. T. Joachims, Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, pp. 169-184, 1999.
9. A. Karlson, G. Robertson, D. Robbins, M. Czerwinski, and G. Smith. FaThumb: a facet-based interface for mobile search. In *Proc. of CHI'06*, pp. 711-720, 2006.
10. Y. Kim and S. Ross. Examining variations of prominent features in genre classification. In *Proc. of Hawaii International Conference on System Sciences*, p. 132. 2008.
11. B. Lee, G. Smith, G. Robertson, M. Czerwinski, D. Tan. FacetLens: exposing trends and relationships to support sensemaking within faceted datasets. In *Proc. of CHI '09*, pp. 1293-1302, 2009.
12. R. Mukherjee and J. Mao. Enterprise search: Tough stuff. *Queue*, pp. 36-46, 2004.
13. M. Plu, L. Agosto, L. Vignollet, J.-C. Marty, A Contact Recommender System for a Mediated Social Media, *Enterprise information systems VI*, Vol. 58, I. Seruca, J. Cordeiro, S. Hammoudi (ed.), Springer, 293-300, 2006.
14. C. Shin and D. S. Doermann. Classification of document page images based on visual similarity of layout structures. In *Proc. SPIE 2000*, pp. 182-190, 2000.
15. H. Simon. *Sciences of the Artificial, 3rd Edition*. MIT Press, Cambridge, Massachusetts, 1996.
16. G. Smith, M. Czerwinski, B. Meyers, D. Robbins, G. Robertson, D. Tan. FacetMap: A scalable search and browse visualization. *IEEE Trans. Visualization and Computer Graphics*, 12, 5, pp. 797-804, 2006.
17. P. Turney. Extraction of Keyphrases from Text: Evaluation of Four Algorithms. *National Research Council of Canada Technical Report ERB-1051*, 1997.
18. M.L. Wilson and m.c. schraefel. A longitudinal study of exploratory and keyword search. In *Proc. of JCDL*, pp. 52-56, 2008.
19. I. Witten, G. Paynter, E. Frank, C. Gutwin, C. Nevill-Manning. KEA: practical automatic keyphrase extraction, In *Proc. of ACM DL*, pp.254-255, 1999.
20. L. Zhen, G. Huang, Z. Jiang, An Inner-Enterprise Knowledge Recommender System, *Expert Systems with Applications*, Elsevier, pp. 1703-1712, 2009.