Exploring the Efficacy of Caption Search for Bioscience Journal Search Interfaces

Marti A. Hearst, Anna Divoli, Jerry Ye
School of Information, UC Berkeley
Berkeley, CA 94720
{hearst,divoli,jerryye}@ischool.berkeley.edu

Michael A. Wooldridge
California Digital Library
Oakland, CA 94612
mikew@ucop.edu

Abstract

This paper presents the results of a pilot usability study of a novel approach to search user interfaces for bioscience journal articles. The main idea is to support search over figure captions explicitly, and show the corresponding figures directly within the search results. Participants in a pilot study expressed surprise at the idea, noting that they had never thought of search in this way. They also reported strong positive reactions to the idea: 7 out of 8 said they would use a search system with this kind of feature, suggesting that this is a promising idea for journal article search.

1 Introduction

For at least two decades, the standard way to search for bioscience journal articles has been to use the National Library of Medicine's PubMed system to search the MEDLINE collection of journal articles. PubMed has innovated search in many ways, but to date search in PubMed is restricted to the title, abstract, and several kinds of metadata about the document, including authors, Medical Subject Heading (MeSH) labels, publication year, and so on.

On the Web, searching within the full text of documents has been standard for more than a decade, and much progress has been made on how to do this well. However, until recently, full text search of bioscience journal articles was not possible due to two major constraints: (1) the full text was not widely available online, and (2) publishers restrict researchers from downloading these articles in bulk.

Recently, online full text of bioscience journal articles has become ubiquitous, eliminating one barrier. The intellectual property restriction is under attack, and we are optimistic that it will be nearly entirely diffused in a few years. In the meantime, the PubMedCentral Open Access collection of journals provides an unrestricted resource for scientists to experiment with for providing full text search.[1]

Full text availability requires a re-thinking of how search should be done on bioscience journal articles. One opportunity is to do information extraction (text mining) to extract facts and relations from the *body* of the text, as well as from the title and abstract as done by much of the early text mining work. (The Biocreative competition includes tasks that allow for extraction within full text (Yeh et al., 2003; Hirschman et al., 2005).) The results of text extraction can then be exposed in search interfaces, as done in systems like iHOP (Hoffmann and Valencia, 2004) and ChiliBot (Chen and Sharp, 2004) (although both of these search only over abstracts).

Another issue is how to adjust search ranking algorithms when using full text journal articles. For example, there is evidence that ranking algorithms should consider which section of an article the query terms are found in, and assign different weights to different sections for different query types (Shah et al., 2003), as seen in the TREC 2006 Genomics Track (Hersh et al., 2006).

Recently Google Scholar has provided search

[1]The license terms for use for BioMed Central can be found at: http://www.biomedcentral.com/info/authors/license and the license for PubMedCentral can be found at: http://www.pubmedcentral.gov/about/openftlist.html

BioNLP 2007: Biological, translational, and clinical language processing, pages 73–80, Prague, June 2007.

over the full text of journal articles from a wide range of fields, but with no special consideration for the needs of bioscience researchers[2]. Google Scholar's distinguishing characteristic is its ability to show the number of papers that cite a given article, and rank papers by this citation count. We believe this is an excellent starting point for full text search, and any future journal article search system should use citation count as a metric. Unfortunately, citation count requires access to the entire collection of articles; something that is currently only available to a search system that has entered into contracts with all of the journal publishers.

In this article, we focus on another new opportunity: the ability to search over figure captions and display the associated figures. This idea is based on the observation, noted by our own group as well as many others, that when reading bioscience articles, researchers tend to start by looking at the title, abstract, figures, and captions. Figure captions can be especially useful for locating information about experimental results. A prominent example of this was seen in the 2002 KDD competition, the goal of which was to find articles that contained experimental evidence for gene products, where the top-performing team focused its analysis on the figure captions (Yeh et al., 2003).

In the Biotext project, we are exploring how to incorporate figures and captions into journal article search explicitly, as part of a larger effort to provide high-quality article search interfaces. This paper reports on the results of a pilot study of the caption search idea. Participants found the idea novel, stimulating, and most expressed a desire to use a search interface that supports caption search and figure display.[3]

2 Related Work

2.1 Automated Caption Analysis

Several research projects have examined the automated analysis of text from captions. Srihari (1991; 1995) did early work on linking information between photographs and their captions, to determine, for example, which person's face in a newspaper photograph corresponded to which name in the caption. Shatkay et al. (2006) combined information from images as well as captions to enhance a text categorization algorithm.

Cohen, Murphy, et al. have explored several different aspects of biological text caption analysis. In one piece of work (Cohen et al., 2003) they devised and tested algorithms for parsing the structure of image captions, which are often quite complex, especially when referring to a figure that has multiple images within it. In another effort, they developed tools to extract information relating to subcellular localization by automatically analyzing fluorescence microscope images of cells (Murphy et al., 2003). They later developed methods to extract facts from the captions referring to these images (Cohen et al., 2003).

Liu et al. (2004) collected a set of figures and classified them according to whether or not they depicted schematic representations of protein interactions. They then allowed users to search for a gene name within the figure caption, returning only those figures that fit within the one class (protein interaction schematics) and contained the gene name.

Yu et al. (2006) created a bioscience image taxonomy (consisting of *Gel-Image*, *Graph*, *Image-of-Thing*, *Mix*, *Model*, and *Table*) and used Support Vector Machines to classify the figures, using properties of both the textual captions and the images.

2.2 Figures in Bioscience Article Search

Some bioscience journal publishers provide a service called "SummaryPlus" that allows for display of figures and captions in the description of a particular article, but the interface does not apply to search results listings.[4]

A medical image retrieval and image annotation task have been part of the ImageCLEF competition since 2005 (Muller et al., 2006).[5] The datasets for this competition are clinical images, and the task is to retrieve images relevant to a query such as "Show blood smears that include polymorphonuclear neu-

[2] http://scholar.google.com

[3] The current version of the interface can be seen at http://biosearch.berkeley.edu

[4] Recently a commercial offering by a company called CSA Illustrata was brought to our attention; it claims to use figures and tables in search in some manner, but detailed information is not freely available.

[5] CLEF stands for Cross-language Evaluation Forum; it originally evaluated multi-lingual information retrieval, but has since broadened its mission.

trophils." Thus, the emphasis is on identifying the content of the images themselves.

Yu and Lee (2006) hypothesized that the information found in the figures of a bioscience article are summarized by sentences from that article's abstract. They succeeded in having 119 scientists mark up the abstract of one of their own articles, indicating which sentence corresponded to each figure in the article. They then developed algorithms to link sentences from the abstract to the figure caption content. They also developed and assessed a user interface called BioEx that makes use of this linking information. The interface shows a set of very small image thumbnails beneath each abstract. When the searcher's mouse hovers over the thumbnail, the corresponding sentence from the abstract is highlighted dynamically.

To evaluate BioEx, Yu and Lee (2006) sent a questionnaire to the 119 biologists who had done the hand-labeling linking abstract sentences to images, asking them to assess three different article display designs. The first design looked like the PubMed abstract view. The second augmented the first view with very small thumbnails of figures extracted from the article. The third was the second view augmented with color highlighting of the abstract's sentences. It is unclear if the biologists were asked to do searches over a collection or were just shown a sample of each view and asked to rate it. 35% of the biologists responded to the survey, and of these, 36 out of 41 (88%) preferred the linked abstract view over the other views. (It should be noted that the effort invested in annotating the abstracts may have affected the scientists' view of the design.)

It is not clear, however, whether biologists would prefer to see the caption text itself rather than the associated information from the abstract. The system described did not allow for searching over text corresponding to the figure caption. The system also did not focus on how to design a full text and caption search system in general.

3 Interface Design and Implementation

The Biotext search engine indexes all Open Access articles available at PubMedCentral. This collection consists of more than 150 journals, 20,000 articles and 80,000 figures. The figures are stored locally, and at different scales, in order to be able to present thumbnails quickly. The Lucene search engine[6] is used to index, retrieve, and rank the text (default statistical ranking). The interface is web-based and is implemented in Python and PHP. Logs and other information are stored and queried using MySQL.

Figure 1a shows the results of searching over the caption text in the Caption Figure view. Figure 1b shows the same search in the Caption Figure with additional Thumbnails (CFT) view. Figure 2a-b shows two examples of the Grid view, in which the query terms are searched for in the captions, and the resulting figures are shown in a grid, along with metadata information.[7] The Grid view may be especially useful for seeing commonalities among topics, such as all the phylogenetic trees that include a given gene, or seeing all images of embryo development of some species.

The next section describes the study participants' reaction to these designs.

4 Pilot Usability Study

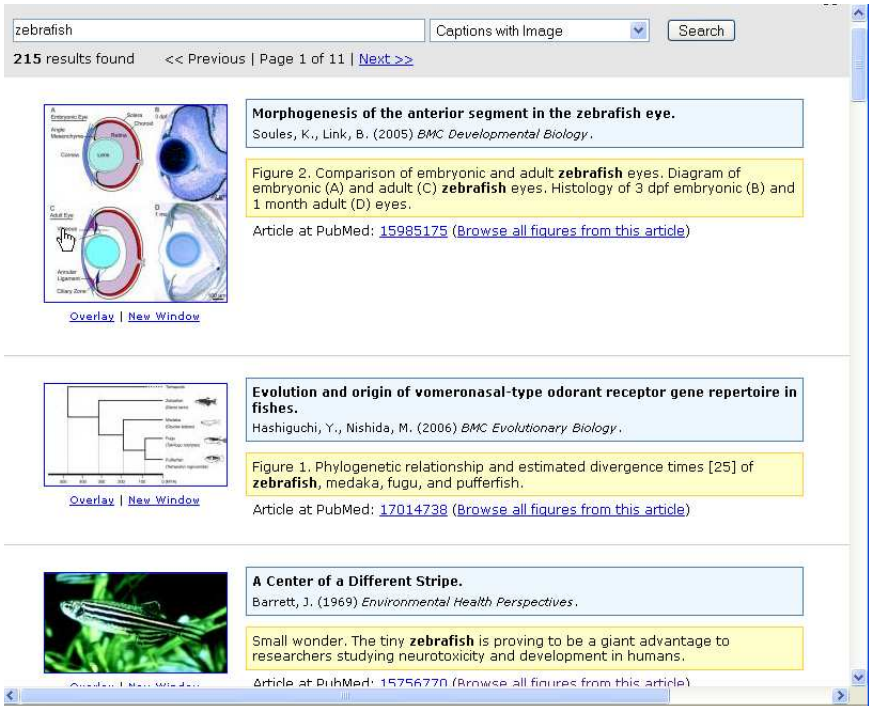
The design of search user interfaces is difficult; the evidence suggests that most searchers are reluctant to switch away from something that is familiar. A search interface needs to offer something qualitatively better than what is currently available in order to be acceptable to a large user base (Hearst, 2006).

Because text search requires the display of text, results listings can quickly obtain an undesirably cluttered look, and so careful attention to detail is required in the elements of layout and graphic design. Small details that users find objectionable can render an interface objectionable, or too difficult to use. Thus, when introducing a new search interface idea, great care must be taken to get the details right. The practice of user-centered design teaches how to achieve this goal: first prototype, then test the results with potential users, then refine the design based on their responses, and repeat (Hix and Hartson, 1993; Shneiderman and Plaisant, 2004).

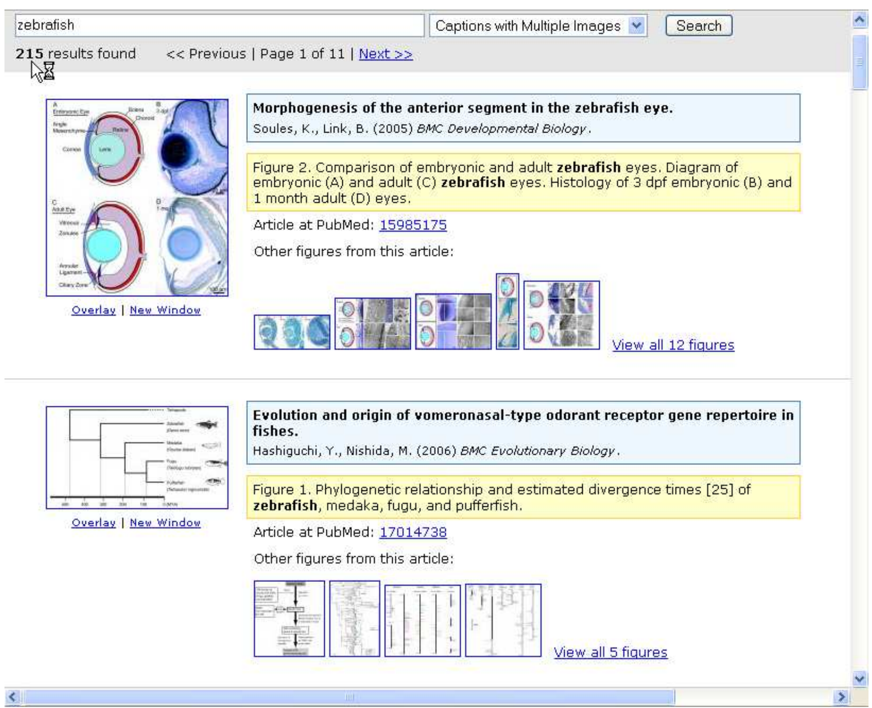
Before embarking on a major usability study to determine if a new search interface idea is a good one, it is advantageous to run a series of pilot studies to determine which aspects of the design work,

[6] http://lucene.apache.org

[7] These screenshots represent the system as it was evaluated. The design has subsequently evolved and changed.



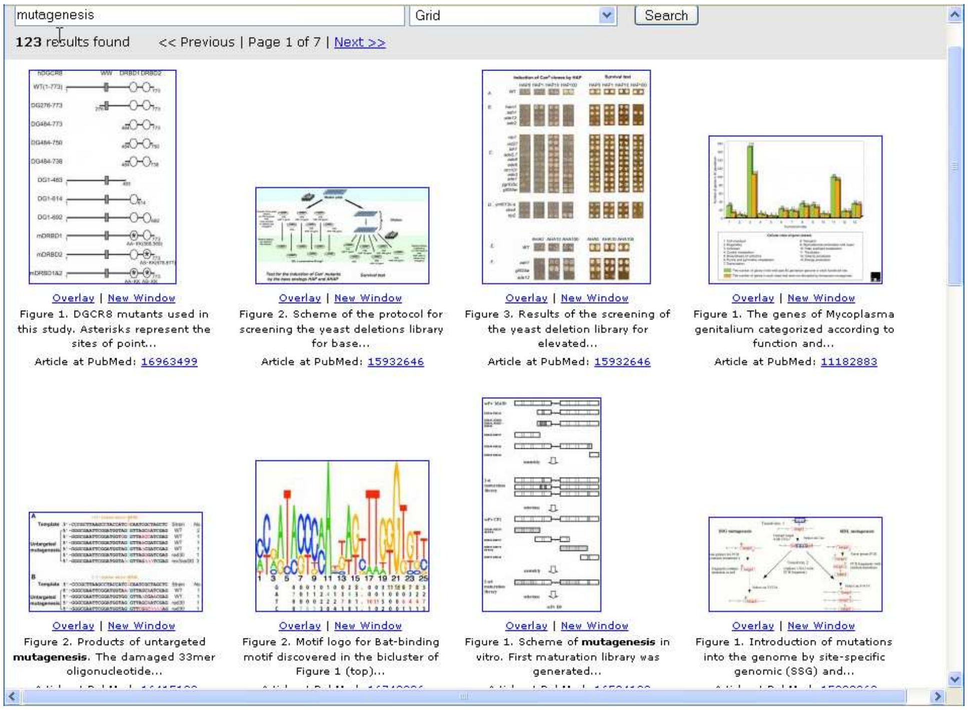
(a)



(b)

Figure 1: Search results on a query of *zebrafish* over the captions within the articles with (a) CF view, and (b) CFT view. The thumbnail is shown to the left of a blue box containing the bibliographic information above a yellow box containing the caption text. The full-size view of the figure can be overlaid over the current page or in a new browser window. In (b) the first few figures are shown as mini-thumbnails in a row below the caption text with a link to view all the figures and captions.

(a)



(b)

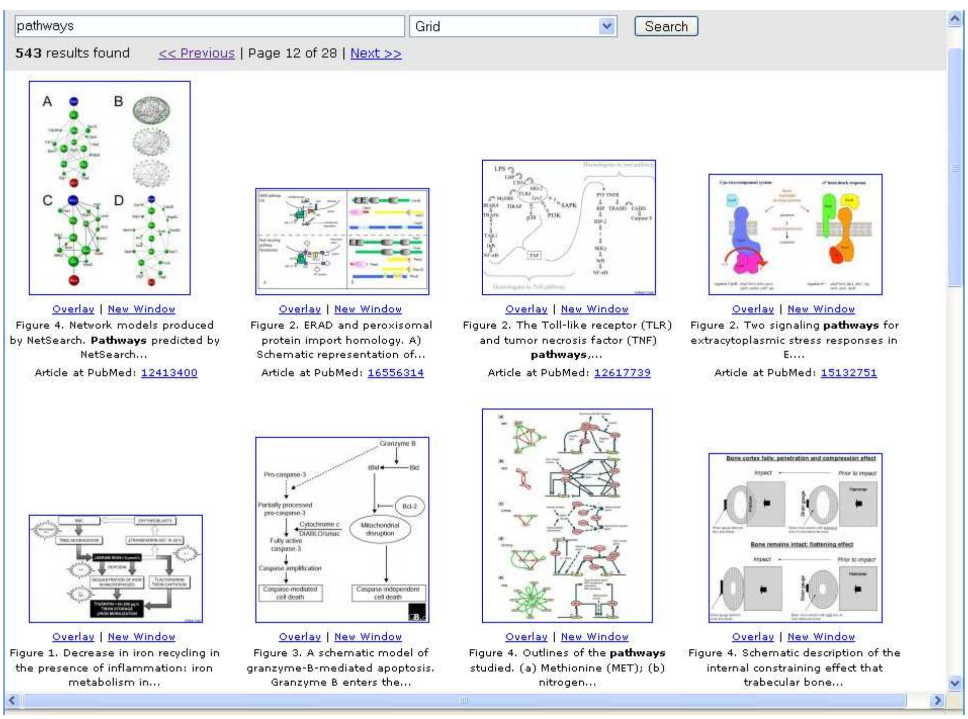


Figure 2: Grid views of the first sets of figures returned as the result of queries for (a) *mutagenesis* and for (b) *pathways* over captions in the Open Access collection.

ID	status	sex	lit search	area(s) of specialization
1	undergrad	F	monthly	organic chemistry
2	graduate	F	weekly	genetics / molecular bio.
3	other	F	rarely	medical diagnostics
4	postdoc	M	weekly	neurobiology, evolution
5	graduate	F	daily	evolutionary bio., entomology
6	undergrad	F	weekly	molecular bio., biochemistry
7	undergrad	F	monthly	cell developmental bio.
8	postdoc	M	daily	molecular / developmental bio.

Table 1: Participant Demographics. Participant 3 is an unemployed former lab worker.

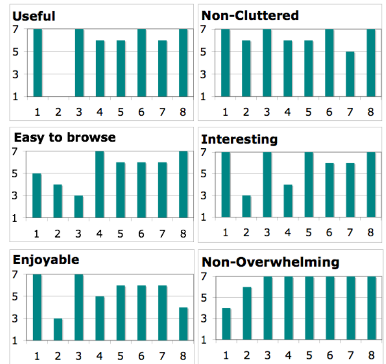


Figure 3: Likert scores on the CF view. X-axis: participant ID, y-axis: Likert scores: 1 = strongly disagree, 7 = strongly agree. (Scale reversed for questionnaire-posed *cluttered* and *overwhelming*.)

which do not, make adjustments, and test some more. Once the design has stabilized and is receiving nearly uniform positive feedback from pilot study participants, then a formal study can be run that compares the novel idea to the state-of-the-art, and evaluates hypotheses about which features work well for which kinds of tasks.

The primary goal of this pilot study was to determine if biological researchers would find the idea of caption search and figure display to be useful or not. The secondary goal was to determine, should caption search and figure display be useful, how best to support these features in the interface. We want to retain those aspects of search interfaces that are both familiar and useful, and to introduce new elements in such a way as to further enhance the search experience without degrading it.

4.1 Method

We recruited participants who work in our campus' main biology buildings to participate in the study. None of the participants were known to us in advance. To help avoid positive bias, we told participants that we were evaluating a search system, but did not mention that our group was the one who was designing the system. The participants all had strong interests in biosciences; their demographics are shown in Table 1.

Each participant's session lasted approximately one hour. First, they were told the purpose of the study, and then filled out an informed consent form and a background questionnaire. Next, they used the search interfaces (the order of presentation was varied). Before the use of each search interface, we explained the idea behind the design. The participant then used the interface to search on their own queries for about 10 minutes, and then filled out a questionnaire describing their reaction to that design. After viewing all of the designs, they filled out a post-study questionnaire where they indicated whether or not they would like to use any of the designs in their work, and compared the design to PubMed-type search.

Along with these standardized questions, we had open discussions with participants about their reactions to each view in terms of design and content. Throughout the study, we asked participants to assume that the new designs would eventually search over the entire contents of PubMed and not just the Open Access collection.

We showed all 8 participants the Caption with Figure (CF) view (see Figure 1a), and Caption with Figure and additional Thumbnails (CFT) (see Figure 1b), as we didn't know if participants would want to see additional figures from the caption's paper.[8] We did not show the first few participants the Grid view, as we did not know how the figure/caption search would be received, and were worried about overwhelming participants with new ideas. (Usability study participants can become frustrated if exposed to too many options that they find distasteful or confusing.) Because the figure search did receive pos-

[8] We also experimented with showing full text search to the first five participants, but as that view was problematic, we discontinued it and substituted a title/abstract search for the remaining three participants. These are not the focus of this study and are not discussed further here.

itive reactions from 3 of the first 4 participants, we decided to show the Grid view to the next 4.

4.2 Results

The idea of caption search and figure display was very positively perceived by all but one participant. 7 out of 8 said they would want to use either CF or CFT in their bioscience journal article searches. Figure 3 shows Likert scores for CF view.

The one participant (number 2) who did not like CF nor CFT thought that the captions/figures would not be useful for their tasks, and preferred seeing the articles' abstracts. Many participants noted that caption search would be better for some tasks than others, where a more standard title & abstract or full-text search would be preferable. Some participants said that different views serve different roles, and they would use more than one view depending on the goal of their search. Several suggested combining abstract and figure captions in the search and/or the display. (Because this could lead to search results that require a lot of scrolling, it would probably be best to use modern Web interface technologies to dynamically expand long abstracts and captions.) When asked for their preference versus PubMed, 5 out of 8 rated at least one of the figure searches above PubMed's interface. (In some cases this may be due to a preference for the layout in our design as opposed to entirely a preference for caption search.)

Two of the participants preferred CFT to CF; the rest thought CFT was too busy. It became clear through the course of this study that it would be best to show all the thumbnails that correspond to a given article as the result of a full-text or abstract-text search interface, and to show only the figure that corresponds to the caption in the caption search view, with a link to view all figures from this article in a new page.

All four participants who saw the Grid view liked it, but noted that the metadata shown was insufficient; if it were changed to include title and other bibliographic data, 2 of the 4 who saw Grid said they would prefer that view over the CF view. Several participants commented that they have used Google Images to search for images but they rarely find what they are looking for. They reacted very positively to the idea of a Google Image-type system specialized to biomedical images. One participant went so far as to open up Google Image search and compare the results directly, finding the caption search to be preferable.

All participants favored the ability to browse all figures from a paper once they find the abstract or one of the figures relevant to their query. Two participants commented that if they were looking for general concepts, abstract search would be more suitable but for a specific method, caption view would be better.

4.3 Suggestions for Redesign

All participants found the design of the new views to be simple and clear. They told us that they generally want information displayed in a simple manner, with as few clicks needed as possible, and with as few distracting links as possible. Only a few additional types of information were suggested from some participants: display, or show links to, related papers and provide a link to the full text PDF directly in the search results, as opposed to having to access the paper via PubMed.

Participants also made clear that they would often want to start from search results based on title and abstract, and then move to figures and captions, and from there to the full article, unless they are doing figure search explicitly. In that case, they want to start with CF or Grid view, depending on how much information they want about the figure at first glance.

They also wished to have the ability to sort the results along different criteria, including year of publication, alphabetically by either journal or author name, and by relevance ranking. This result has been seen in studies of other kinds of search interfaces as well (Reiterer et al., 2005; Dumais et al., 2003). We have also received several requests for table caption search along with figure caption search.

5 Conclusions and Future Work

The results of this pilot study suggest that caption search and figure display is a very promising direction for bioscience journal article search, especially paired with title/abstract search and potentially with other forms of full-text search. A much larger-scale study must be performed to firmly establish this result, but this pilot study provides insight about how to design a search interface that will be positively received in such a study. Our results also suggest that web search systems like Google Scholar and Google Images could be improved by showing images from the articles along lines of specialization.

The Grid view should be able to show images grouped by category type that is of interest to biologists, such as heat maps and phylogenetic trees. One participant searched on *pancreas* and was surprised when the top-ranked figure was an image of a machine. This idea underscores the need for BioNLP research in the study of automated caption classification. NLP is needed both to classify images and perhaps also to automatically determine which images are most "interesting" for a given article.

To this end, we are in the process of building a classifier for the figure captions, in order to allow for grouping by type. We have developed an image annotation interface and are soliciting help with hand-labeling from the research community, to build a training set for an automated caption classifier.

In future, we plan to integrate table caption search, to index the text that refers to the caption, along with the caption, and to provide interface features that allow searchers to organize and filter search results according to metadata such as year published, and topical information such as genes/proteins mentioned. We also plan to conduct formal interface evaluation studies, including comparing to PubMed-style presentations.

Acknowledgements: This work was supported in part by NSF DBI-0317510. We thank the study participants for their invaluable help.

References

H. Chen and B.M. Sharp. 2004. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5(147).

W.W. Cohen, R. Wang, and R.F. Murphy. 2003. Understanding captions in biomedical publications. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 499–504.

S. Dumais, E. Cutrell, J.J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. *Proceedings of SIGIR 2003*, pages 72–79.

M. Hearst. 2006. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR Workshop on Faceted Search*, Seattle, WA.

W. Hersh, A. Cohen, P. Roberts, and Rekapalli H. K. 2006. TREC 2006 genomics track overview. *The Fifteenth Text Retrieval Conference*.

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6:1.

D. Hix and H.R. Hartson. 1993. *Developing user interfaces: ensuring usability through product & process*. John Wiley & Sons, Inc. New York, NY, USA.

R. Hoffmann and A. Valencia. 2004. A gene network for navigating the literature. *Nature Genetics*, 36(664).

F. Liu, T-K. Jenssen, V. Nygaard, J. Sack, and E. Hovig. 2004. FigSearch: a figure legend indexing and classification system. *Bioinformatics*, 20(16):2880–2882.

H. Muller, T. Deselaers, T. Lehmann, P. Clough, E. Kim, and W. Hersh. 2006. Overview of the ImageCLEF 2006 Medical Image Retrieval Tasks. In *Working Notes for the CLEF 2006 Workshop*.

R.F. Murphy, M. Velliste, and G. Porreca. 2003. Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images. *The Journal of VLSI Signal Processing*, 35(3):311–321.

B. Rafkind, M. Lee, S.F. Chang, and H. Yu. 2006. Exploring text and image features to classify images in bioscience literature. *Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL*, 6:73–80.

H. Reiterer, G. Tullius, and T. M. Mann. 2005. Insyder: a content-based visual-information-seeking system for the web. *International Journal on Digital Libraries*, 5(1):25–41, Mar.

P.K. Shah, C. Perez-Iratxeta, P. Bork, and M.A. Andrade. 2003. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4(20).

H. Shatkay, N. Chen, and D. Blostein. 2006. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446.

B. Shneiderman and C. Plaisant. 2004. *Designing the user interface: strategies for effective human-computer interaction, 4/E*. Addison Wesley.

R.K. Srihari. 1991. PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs. *Proceedings AAAI-91*, pages 80–85.

RK Srihari. 1995. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56.

A.S. Yeh, L. Hirschman, and A.A. Morgan. 2003. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19(1):i331–i339.

H. Yu and M. Lee. 2006. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22(14):e547.