BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 17 2008, pages 1968–1970
doi:10.1093/bioinformatics/btn340

Data and text mining

Yale Image Finder (YIF): a new search engine for retrieving biomedical images

Songhua Xu[1], James McCusker[2] and Michael Krauthammer[2,*]

[1]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06520 and [2]Department of Pathology & Yale Center for Medical Informatics, 300 Cedar Street, New Haven, CT 06510, USA

Received on February 4, 2008; revised on June 9, 2008; accepted on July 2, 2008

Advance Access publication July 9, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Yale Image Finder (YIF) is a publicly accessible search engine featuring a new way of retrieving biomedical images and associated papers based on the text carried inside the images. Image queries can also be issued against the image caption, as well as words in the associated paper abstract and title. A typical search scenario using YIF is as follows: a user provides few search keywords and the most relevant images are returned and presented in the form of thumbnails. Users can click on the image of interest to retrieve the high resolution image. In addition, the search engine will provide two types of related images: those that appear in the same paper, and those from other papers with similar image content. Retrieved images link back to their source papers, allowing users to find related papers starting with an image of interest. Currently, YIF has indexed over 140 000 images from over 34 000 open access biomedical journal papers.

Availability: http://krauthammerlab.med.yale.edu/imagefinder/

Contact: michael.krauthammer@yale.edu

1 INTRODUCTION AND RELATED WORK

With the steady increase of publications in biomedicine, it is getting ever more difficult to stay on top of the latest research results. Web-based information retrieval engines, such as Google and Yahoo, are key for navigating biomedical documents posted on the web, while other search engines, such as Entrez, are essential in locating documents that are stored and indexed in domain databases such as PubMed. There is ongoing research and development in building tailored search engines for finding biomedical research papers, as exemplified by research done in context of the TREC challenges (Cohen and Hersh, 2006).

Several teams have recently presented image-based systems and methodologies for facilitating the information retrieval process. The BioText project has built a search engine that allows for searches over image captions (Hearst *et al*., 2007). Qian and Murphy (2008) describe a system for accessing fluorescence microscopy images via image classification and segmentation. Also, Shatkay *et al*. (2006) have proposed to incorporate image data for text categorization. Most recently, Jing and Baluja (2008) modified the conventional Google PageRank algorithm for image search based on image similarities estimated from low-level visual features.

*To whom correspondence should be addressed.

However, we are not aware of a biomedical search engine that can retrieve images by searching the text within biomedical images. This offers several advantages over searching over captions alone. First, captions may not contain all the textual information that is contained in the images. Second, image texts are usually very specific, allowing for precise matching of images with related images. Here, we discuss Yale Image Finder (YIF), which allows for querying for images over image text, image captions, as well as abstracts and titles of the associated papers.

2 THE USER VIEW OF THE SEARCH ENGINE

Interface for submitting queries based on keywords A user can provide a few keywords to form a query, which can be formulated using Boolean operators. Via a checkbox, he can restrict the queries to the text within the images, the image caption, the paper title, paper abstract, full text or any combination thereof. An example query is shown in Figure 1.

Interface for the thumbnail view Once a query is submitted, all the retrieved images will be returned and presented in a thumbnail view with an image caption excerpt, see Figure 1.

Interface for viewing an image in high resolution The image thumbnail links to a page where a high-resolution version of the image is presented. The recognized image text, the caption of the image, its paper's title and abstract, and the link to the original paper are provided. A special feature is the display of related images on the bottom and right sides of the page. On the bottom are the thumbnails of images that were published in the same paper. On the right are related images across all documents in our database, where relatedness is primarily determined by the similarity of the words within images.

3 IMPLEMENTATION

Right now, YIF indexes over 140 000 images from over 34 000 open access papers from PubMed Central. The system is updated on a regular basis. The key idea in our newly proposed technology is that we provide customized layout analysis over images published in academic journals, using histogram-based image processing techniques (Manmatha and Riseman, 1999). The analysis identifies image text elements, and subjects them to optical character recognition (OCR). The text extraction is repeated after turning an

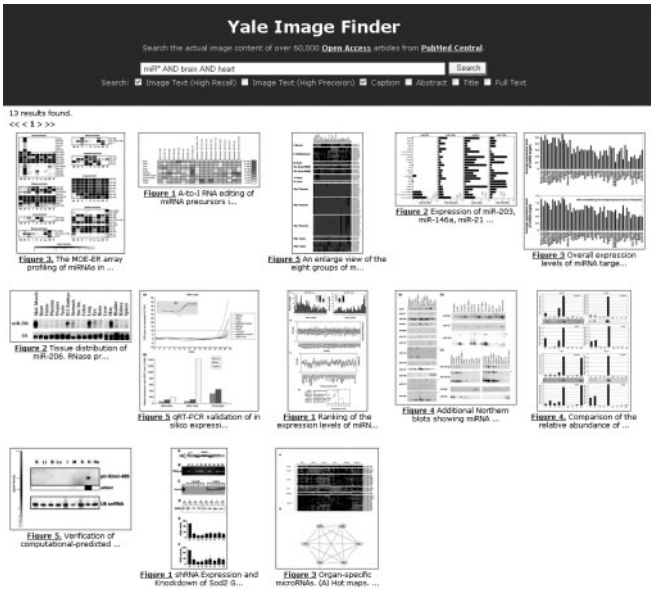

Fig. 1. Initial search result page after a user submits a search query. The query shown is Query No. 3 in Table 2.

Table 1. Performance statistics for OCR results over 161 randomly selected images

	High-recall mode	High-precision mode
Precision	27.85%	87.68%
Recall	64.79%	38.45%
F-rate	0.390	0.535

Manual labeling determined that the images contain 2445 text strings, of which 70.84% are not found in the associated captions.

image 90°, to allow for the capture of vertical image labels. In order to minimize false positive results, we optionally perform a cross-checking procedure of the extracted image text against the full text of the articles. We only retain image text that is mentioned in the articles (including image captions), assuming that articles usually discuss the content of their images. We thus process image text in two ways. Once by subjecting image text to the cross-checking procedure ('high-precision mode'), and once by skipping the procedure ('high-recall mode'). We then index the images and the extracted text with Apache Lucene, an Open Source search engine library (Cutting *et al*., 2008).

We conducted an evaluation study to determine the accuracy of text extraction. We first generated an image corpus of 161 randomly selected images. The images were part of the open access image collection from PubMed Central. We then manually wrote out all the strings appearing in those images, excluding strings consisting of numbers or symbols only. This resulted in a corpus of 2445 image text strings. We then compared the automatically extracted with the manually extracted strings, and generated the following statistics: text extraction recall, precision and F-score for the high-recall and high-precision modes (Table 1). Our system retrieves 64.79% of the actual image text content at 27.85% precision, in the high-recall mode, and 38.45% of the image text content at 87.68% precision, in the high-precision mode.

In order to assess the actual image retrieval performance, we conducted an additional evaluation using three typical image queries (Table 2). Compared to searches that are restricted to the image caption alone, we found that our search engine retrieves additional images, particularly of types 'diagram' and 'list' (i.e. lists of GO terms, but also lists of genes as featured in heatmap images). This is intuitively understandable, as authors often do not mention all the elements from diagrams or list-type images in the associated captions. For both queries #1 and #2, querying the image text retrieved ∼30% additional images showing the relationship between 'diet' and 'insulin', and 'p53' and 'apoptosis', respectively. For query #3, querying the image text more than doubled the number of images showing miRNA expression across different cell types, indicating that authors consistently place specific information (such as the names of tissues) in the image itself, rather than the caption.

Finally, in all our three queries reported in Table 2, the precision of searching against caption and image text is high (> 80%), indicating that the low precision of the OCR procedure itself only modestly affects the performance of actual image queries. The reason is that many of the wrongly recognized image strings (OCR errors) are

Table 2. Performance statistics for three image queries

No.	Query	Search target domain	Graph	Gel	Microscopy	Diagram	List	Misc.	Total	Relevant images	Precision
1	diet AND insulin	Caption	17	0	0	0	0	2	19	19	100%
		Caption+Image Text (HR)	17	1	0	6	0	3	27	25	92.59%
		Δ	0	1	0	6	0	1	8 (42.11%)	6 (31.58%)	–
2	apoptosis AND p53	Caption	11	1	5	11	0	14	42	42	100%
		Caption+Image Text (HR)	12	1	5	18	4	15	55	54	98.18%
		Δ	1	0	0	7	4	1	13 (30.95%)	12 (28.57%)	–
3	miR*	Caption	1	1	0	0	1	2	5	4	80%
	AND brain AND heart	Caption+Image Text (HR)	1	3	0	0	6	3	13	11	84.62%
		Δ	0	2	0	0	5	1	8 (160%)	7 (175%)	–

The second column lists the actual queries entered into YIF; the third column specifies the search target domain, where 'Image Text (HR)' stands for 'Image Text (High Recall)'. The row titled 'Δ' indicates the number of additional images found by using the 'Caption+Image Text (HR)' option versus the 'Caption' option, i.e. the additional images found by querying against image text. The fourth to the 10th columns show the number of images retrieved broken down according to image categories, as well as the total number of images found. The 11th column lists the number of retrieved and relevant images, as judged by a human expert, and the 12th column indicates overall image search precision. In the third query, the asterisk in miR* is useful for retrieving different types and spelling variations of miRNAs, such as mir-1 and mir-22.

non-sensical, and will never be used in an actual image query. Additional experiments (data not shown) indicate that the precision of image queries drops for very short search strings. We recommend to use the 'high precision' mode for such types of queries.

4 DISCUSSION

We present YIF, a novel search engine that indexes text found inside biomedical images. YIF offers more comprehensive research results by searching over text that may not be present in the image caption, and offers the ability to find related images and associated papers by directly comparing image content. We believe that searching over image text opens up new avenues for fruitful research in biomedical information retrieval.

ACKNOWLEDGEMENTS

Funding: This research has been funded by NLM grant 5K22LM009255.

Conflict of Interest: none declared.

REFERENCES

Hearst,M. *et al*. (2007) BioText Search Engine: beyond abstract search. *Bioinformatics*, **23**, 2196–2197.

Qian,Y. and Murphy,R.F. (2008) Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. *Bioinformatics*, **24**, 569–576.

Cohen,A.M. and Hersh,W.R. (2006) The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J. Biomed. Discov. Collab.*. **1**, 4.

Shatkay,H. *et al*. (2006) Integrating image data into biomedical text categorization. *Bioinformatics*, **22**, e446–e453.

Cutting,D. *et al*. (2008) Apache Lucene. Available at http://lucene.apache.org/java/docs/ (last accessed date 22 July, 2008).

Manmatha,V.W. and Riseman,E.M. (1999) Textfinder: an automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Anal. Mach. Intell.*, **21**, pp. 1224–1229.

Jing,Y. and Baluja,S. (2008) PageRank for product image search. *WWW 2008: Proceedings of the 17th International World Wide Web Conference*, Beijing, China. ACM Press, New York, NY, USA, pp. 307–315.