BioEx: A Novel User-Interface that Accesses Images from Abstract Sentences

Hong Yu
Department of Biomedical Informatics
Columbia University
New York, NY 10032
Hy52@columbia.edu

Minsuk Lee
Department of Biomedical Informatics
Columbia University
New York, NY 10032
minsuk.lee@gmail.com

Abstract

Images (i.e., figures or tables) are important experimental results that are typically reported in bioscience full-text articles. Biologists need to access the images to validate research facts and to formulate or to test novel research hypotheses. We designed, evaluated, and implemented a novel user-interface, BioEx, that allows biologists to access images that appear in a full-text article directly from the abstract of the article.

1 Introduction

The rapid growth of full-text electronic publications in bioscience has made it necessary to create information systems that allow biologists to navigate and search efficiently among them. Images are usually important experimental results that are typically reported in full-text bioscience articles. An image is worth a thousand words. Biologists need to access image data to validate research facts and to formulate or to test novel research hypotheses. Additionally, full-text articles are frequently long and typically incorporate multiple images. For example, we have found an average of 5.2 images per biological article in the journal *Proceedings of the National Academy of Sciences* (PNAS). Biologists need to spend significant amount of time to read the full-text articles in order to access specific images.

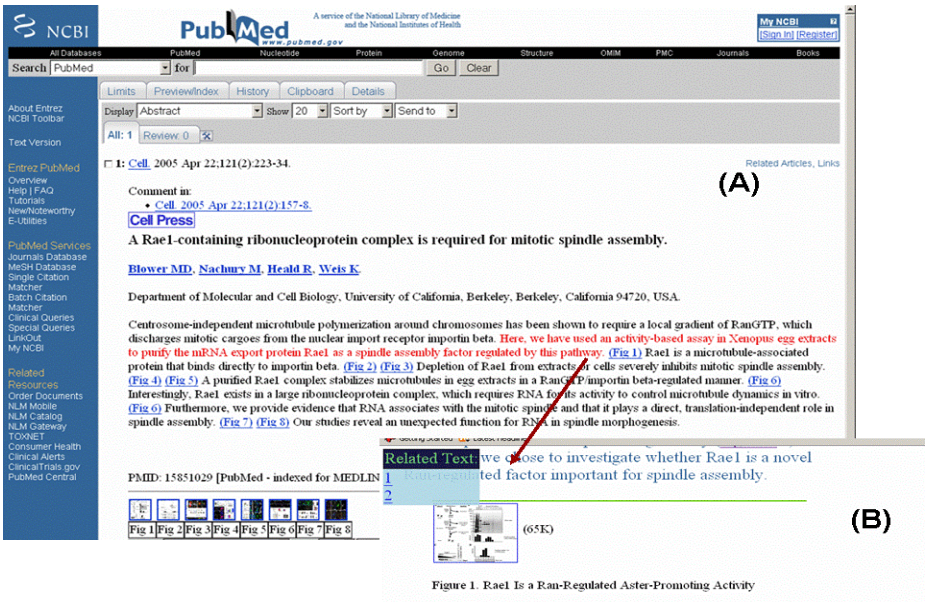

Figure 1. BioEx user-interface (as shown in A) is built upon the PubMed user-interface. Images are shown as thumbnails at the bottom of a PubMed abstract. Images include both Figure and Table. When a mouse (as shown as a hand in A) moves to "Fig x", it shows the associated abstract sentence(s) that link to the original figure that appears in the full-text articles. For example, "Fig 1" links to image B. "Related Text" provides links to other associated texts that correspond to the image besides its image caption.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 189–192,
New York, June 2006.

In order to facilitate biologists' access to images, we designed, evaluated, and implemented a novel user-interface, BioEx, that allows biologists to access images that appear in a full-text article directly from the abstract of the article. In the following, we will describe the BioEx user-interface, evaluation, and the implementation.

2. Data Collection

We hypothesize that images reported in a full-text article can be summarized by sentences in the abstract. To test this hypothesis, we randomly selected a total of 329 biological articles that are recently published in leading journals *Cell* (104), *EMBO* (72), *Journal of Biological Chemistry* (92), and *Proceedings of the National Academy of Sciences (PNAS)* (61). For each article, we e-mailed the corresponding author and invited him or her to identify abstract sentences that summarize image content in that article. In order to eliminate the errors that may be introduced by sentence boundary ambiguity, we manually segmented the abstracts into sentences and sent the sentences as the email attachments.

A total of 119 biologists from 19 countries participated voluntarily the annotation to identify abstract sentences that summarize figures or tables from 114 articles (39 *Cells*, 29 *EMBO*, 30 *Journal of Biological Chemistry*, and 16 *PNAS*), a collection that is 34.7% of the total articles we requested. The responding biologists included the corresponding authors to whom we had sent emails, as well as the first authors of the articles to whom the corresponding authors had forwarded our emails. None of the biologists or authors were compensated.

This collection of 114 full-text articles incorporates 742 images and 826 abstract sentences. The average number of images per document is 6.5±1.5 and the average number of sentences per abstract is 7.2±1.9. Our data show that 87.9% images correspond to abstract sentences and 66.5% of the abstract sentences correspond to images. The data empirically validate our hypothesis that image content can be summarized by abstract sentences. Since an abstract is a summary of a full-text article, our results also empirically validate that images are important elements in full-text articles. This collection of 114 annotated articles was then used as the corpus to evaluate automatic mapping of abstract sentences to images using the natural language processing approaches described in Section 4.

3. BioEx User-Interface Evaluation

In order to evaluate whether biologists would prefer to accessing images from abstract sentence links, we designed BioEx (Figure 1) and two other baseline user-interfaces. BioEx is built upon the PubMed user-interface except that images can be accessed by the abstract sentences. We chose the PubMed user-interface because it has more than 70 million hits a month and represents the most familiar user-interface to biologists. Other information systems have also adapted the PubMed user-interface for similar reasons (Smalheiser and Swanson 1998; Hearst 2003). The two other baseline user-interfaces were the original PubMed user-interface and a modified version of the SummaryPlus user-interface, in which the images are listed as disjointed thumbnails rather than related by abstract sentences.

We asked the 119 biologists who linked sentences to images in their publications to assign a label to each of the three user-interfaces to be "My favorite", "My second favorite", or "My least favorite". We designed the evaluation so that a user-interface's label is independent of the choices of the other two user-interfaces.

A total of 41 or 34.5% of the biologists completed the evaluation in which 36 or 87.8% of the total 41 biologists judged BioEx as "My favorite". One biologist judged all three user-interfaces to be "My favorite". Five other biologists considered SummaryPlus as "My favorite", two of whom (or 4.9% of the total 41 biologists) judged BioEx to be "My least favorite".

4. Linking Abstract Sentences to Images

We have explored hierarchical clustering algorithms to cluster abstract sentences and image captions based on lexical similarities.
Hierarchical clustering algorithms are well-established algorithms that are widely used in

many other research areas including biological sequence alignment (Corpet 1988), gene expression analyses (Herrero et al. 2001), and topic detection (Lee et al. 2006). The algorithm starts with a set of text (i.e., abstract sentences or image captions). Each sentence or image caption represents a document that needs to be clustered. The algorithm identifies pair-wise document similarity based on the TF*IDF weighted cosine similarity. It then merges the two documents with the highest similarity into one cluster. It then re-evaluates pairs of documents/clusters; two clusters can be merged if the average similarity across all pairs of documents within the two clusters exceeds a predefined threshold. In presence of multiple clusters that can be merged at any time, the pair of clusters with the highest similarity is always preferred.

In our application, if abstract sentences belong to the same cluster that includes images captions, the abstract sentences summarize the image content of the corresponded images. The clustering model is advantageous over other models in that the flexibility of clustering methods allows "many-to-many" mappings. That is a sentence in the abstract can be mapped to zero, one or more than one images and an image can be mapped to zero, one or more than one abstract sentences.

We explored different learning features, weights and clustering algorithms to link abstract sentences to images. We applied the TF*IDF weighted cosine similarity for document clustering. We treat each sentence or image caption as a "document" and the features are bag-of-words.

We tested three different methods to obtain the IDF value for each word feature: 1) ***IDF(abstract+caption):*** the IDF values were calculated from the pool of abstract sentences and image captions; 2) ***IDF(full-text):*** the IDF values were calculated from all sentences in the full-text article; and 3) ***IDF(abstract)::IDF(caption):*** two sets of IDF values were obtained. For word features that appear in abstracts, the IDF values were calculated from the abstract sentences. For words that appear in image captions, the IDF values were calculated from the image captions.

The positions of abstract sentences or images are important. The chance that two abstract sentences link to an image decreases when the distance between two abstract sentences increases. For example, two consecutive abstract sentences have a higher probability to link to one image than two abstract sentences that are far apart. Two consecutive images have a higher chance to link to the same abstract sentence than two images that are separated by many other images. Additionally, sentence positions in an abstract seem to correspond to image positions. For example, the first sentences in an abstract have higher probabilities than the last sentences to link to the first image.

To integrate such "neighboring effect" into our existing hierarchical clustering algorithms, we modified the TF*IDF weighted cosine similarity. The TF*IDF weighted cosine similarity for a pair of documents *i* and *j* is *Sim(i,j)*, and the final similarity metric *W(i,j)* is:

$$W(i, j) = Sim(i, j) * (1 - abs(P_i / T_i - P_j / T_j))$$

1. If *i* and *j* are both abstract sentences, $T_i=T_j$*=total number of abstract sentences;* and P_i and P_j represents the positions of sentences *i* and *j* in the abstract.

2. If *i* and *j* are both image captions, $T_i=T_j$*=total number of images that appear in a full-text article;* and P_i and P_j represents the positions of images *i* and *j* in the full-text article.

3. If *i* and *j* are an abstract sentence and an image caption, respectively, T_i*=total number of abstract sentences* and T_j*=total number of images that appear in a full-text article;* and P_i and P_j represent the positions of abstract sentence *i* and image *j*.

Finally, we explored three clustering strategies; namely, *per-image, per-abstract sentence*, and *mix.*

The **Per-image** strategy clusters each image caption with all abstract sentences. The image is

assigned to (an) abstract sentence(s) if it belongs to the same cluster. This method values features in abstract sentences more than image captions because the decision that an image belongs to (a) sentence(s) depends upon the features from all abstract sentences and the examined image caption. The features from other image captions do not play a role in the clustering methodology.

The **Per-abstract-sentence** strategy takes each abstract sentence and clusters it with all image captions that appear in a full-text article. Images are assigned to the sentence if they belong to the same cluster. This method values features in image captions higher than the features in abstract sentences because the decision that an abstract sentence belongs to image(s) depends upon the features from the image captions and the examined abstract sentence. Similar to per-image clustering, the features from other abstract sentences do not play a role in the clustering methodology.

The **Mix** strategy clusters all image captions with all abstract sentences. This method treats features in abstract sentences and image captions equally.

5. Results and Conclusions

Figures 2 - 4 show the results from three different combinations of features and algorithms with varied TF*IDF thresholds. The default parameters for all these experiments were "per image", "bag-of-words", and "without neighboring weight".

Figure 2 shows that the "global" IDFs, or the IDFs obtained from the full-text article, have a much lower performance than "local" IDFs, or IDFs calculated from the abstract sentences and image captions. Figure 3 shows that **Per-image** out-performs the other two strategies. The results suggest that features in abstract sentences are more useful than features that reside within captions for the task of clustering. Figure 4 shows that the "neighboring weighted" approach offers significant enhancement over the TF*IDF weighted approach. When the recall is 33%, the precision of "neighboring weighted" approach increases to 72% from the original 38%, which corresponds to a 34% increase. The results strongly indicate the importance of the "neighboring effect" or positions of additional features. When the precision is 100%, the recall is 4.6%. We believe BioEx system is applicable for real use because a high level of precision is the key to BioEx success.

Acknowledgement: The authors thank Dr. Weiqing Wang for her contribution to this work. The authors also thank Michael Bales, Li Zhou and Eric Silfen, and three anonymous reviewers for valuable comments. The authors acknowledge the support of Juvenile Diabetes Foundation International (JDRF 6-2005-835).

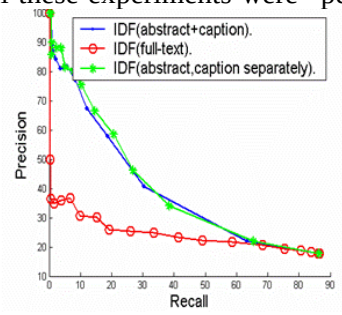

Figure 2

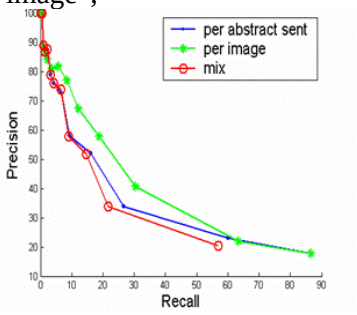

Figure 3

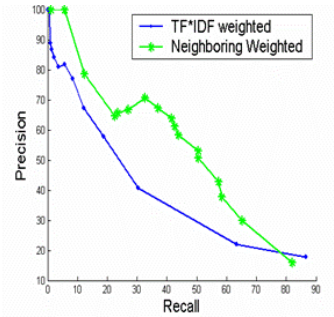

Figure 4

References:

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881-10890

Hearst M (2003) The BioText project. A powerpoint presentation.

Herrero J, Valencia A, Dopazo J (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 17:126-136

Lee M, Wang W, Yu H (2006) Exploring supervised and unsupervised methods to detect topics in Biomedical text. BMC Bioinformatics 7:140

Smalheiser NR, Swanson DR (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed 57:149-153