BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 12 2011, pages 1723–1724
doi:10.1093/bioinformatics/btr194

Data and text mining
Advance Access publication April 14, 2011

Figure summarizer browser extensions for PubMed Central

Shashank Agarwal[1,*] and Hong Yu[1,2,3,*]

[1]Medical Informatics, [2]Department of Computer Science and Electrical Engineering and [3]Department of Health Sciences, College of Health Science, University of Wisconsin-Milwaukee, 2200 E. Kenwood Blvd., Milwaukee WI 53201-0413, USA

Associate Editor: John Quackenbush

ABSTRACT

Summary: Figures in biomedical articles present visual evidence for research facts and help readers understand the article better. However, when figures are taken out of context, it is difficult to understand their content. We developed a summarization algorithm to summarize the content of figures and used it in our figure search engine (http://figuresearch.askhermes.org/). In this article, we report on the development of web browser extensions for Mozilla Firefox, Google Chrome and Apple Safari to display summaries for figures in PubMed Central and NCBI Images.

Availability: The extensions can be downloaded from http://figuresearch.askhermes.org/articlesearch/extensions.php.

Contact: agarwal@uwm.edu

Received on January 12, 2011; revised on March 1, 2011; accepted on April 6, 2011

1 INTRODUCTION

Biomedical scientists need to access figures to validate research facts and to formulate or test novel research hypotheses. However, figures are difficult to comprehend without associated text (e.g. figure legend and other reference text). Our evaluation has shown statistically significant differences in figure comprehension when varying levels of text were provided (Yu *et al.*, 2009). When the full-text article is not available, presenting just the figure+legend left biomedical researchers lacking 39–68% of the information about a figure as compared to having complete figure comprehension; adding the title and abstract improved the situation, but still left biomedical researchers missing 30% of the information. When the full-text article is available, figure comprehension increased to the highest 86–97% (Yu *et al.*, 2009). The results indicate that there is information in the abstract and in the full-text that biomedical scientists require to understand the figures that appear in biomedical articles (Yu *et al.*, 2009). On the other hand, we also found that the associated text of a figure is typically distributed across the full-text body and is frequently redundant in content (Yu, 2006). For example, the following three redundant sentences in the abstract, introduction, and caption describe an image (Figure 1) of the article (Das *et al.*, 2003):

> Abstract: *PTEN is composed of an N-terminal phosphatase domain, a C2 domain, and a C-terminal tail region that contains the PSD-95/Dlg/ZO-1 homology (PDZ) domain-binding sequence and multiple phosphorylation sites.*
>
> Introduction: *PTEN is composed of the N-terminal phosphatase domain (≈ 180 aa), the C2 domain (≈ 165 aa), and the C-terminal tail (≈ 50 aa) (see Fig. 1).*
>
> Figure Caption: *PTEN has a N-terminal phosphatase domain, a C2 domain, and a C-terminal tail that contains multiple phosphorylation sites and a PDZ domain-binding sequence.*

We hypothesize that a succinct and structured summary will allow biomedical researchers to comprehend the figure data efficiently and to access relevant figures in a timely manner. Figure summaries may benefit other text-mining tasks, including information extraction, retrieval and question answering. For example, information extraction systems may target fewer and more succinct statements. Information retrieval and question answering can benefit by removing redundant information.

We developed methods (Agarwal and Yu, 2009a, b; Yu *et al.*, 2009) to automatically aggregate distributed associated text, remove redundant information and generate a text summary for each figure. We first found that biomedical researchers prefer a structured summary that follows the IMRaD format (Introduction, Methods, Results and Discussion) (Yu *et al.*, 2009). We developed a supervised machine learning classifier to automatically classify sentences appearing in a full-text biomedical articles into the IMRaD format with an F1-score of 91.55% (Agarwal and Yu, 2009a). We then developed a simple information retrieval approach that selects one sentence from each of the IMRaD categories based on cosine similarity between the sentence and the image caption (Agarwal and Yu, 2009b). We generated a manual gold-standard for 44 figures in 7 articles by asking annotators to select sentences from the article that best summarize the figure. We then evaluated our figure summarizer using the ROUGE score (recall-oriented understudy for gisting evaluation) (Lin, 2004) and obtained an average ROUGE-1 score of 0.70. In comparison, the abstract of the article attained a ROUGE-1 score of 0.33. In the future, we plan to conduct a cognitive evaluation as we have done in (Yu *et al.*, 2009) to evaluate how automatically generated figure summaries improve figure comprehension. We implemented the figure summarization system with over 200 000 full-text biomedical articles deposited in PubMed Central's Open Access Subset as an online system called 'figuresearch' (http://figuresearch.askhermes.org). We conducted an online survey with 24 participants, and 65.2% survey participants found that the summaries are useful for figure comprehension.

In October 2010, the National Library of Medicine (NLM) announced that PubMed Abstract display for PubMed Central® articles are enhanced to include an image strip generated from

[*]To whom correspondence should be addressed.

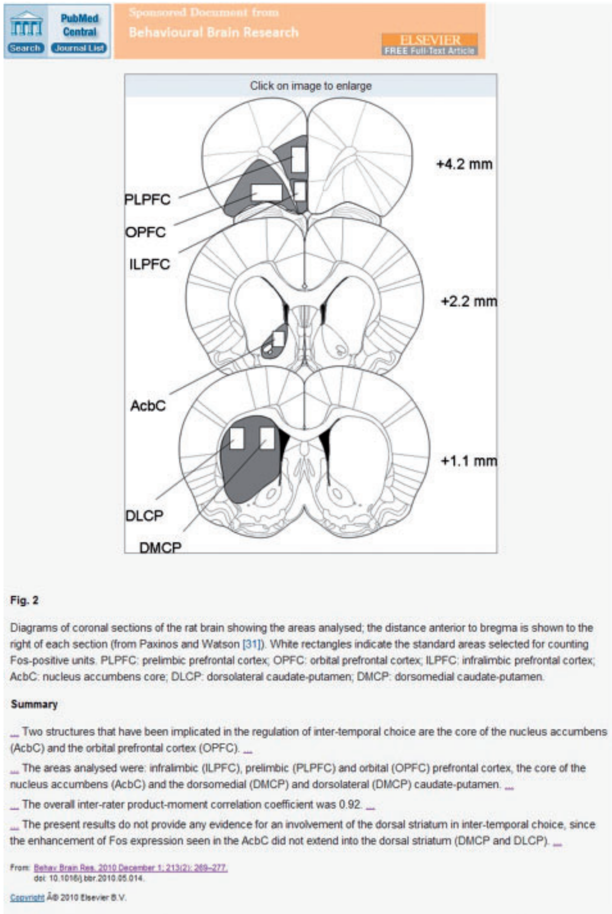

Fig. 2

Diagrams of coronal sections of the rat brain showing the areas analysed; the distance anterior to bregma is shown to the right of each section (from Paxinos and Watson [31]). White rectangles indicate the standard areas selected for counting Fos-positive units. PLPFC: prelimbic prefrontal cortex; OPFC: orbital prefrontal cortex; ILPFC: infralimbic prefrontal cortex; AcbC: nucleus accumbens core; DLCP: dorsolateral caudate-putamen; DMCP: dorsomedial caudate-putamen.

Summary

_ Two structures that have been implicated in the regulation of inter-temporal choice are the core of the nucleus accumbens (AcbC) and the orbital prefrontal cortex (OPFC). _

_ The areas analysed were: infralimbic (ILPFC), prelimbic (PLPFC) and orbital (OPFC) prefrontal cortex, the core of the nucleus accumbens (AcbC) and the dorsomedial (DMCP) and dorsolateral (DMCP) caudate-putamen. _

_ The overall inter-rater product-moment correlation coefficient was 0.92. _

_ The present results do not provide any evidence for an involvement of the dorsal striatum in inter-temporal choice, since the enhancement of Fos expression seen in the AcbC did not extend into the dorsal striatum (DMCP and DLCP). _

From: Behav Brain Res. 2010 December 1; 213(2): 269–277.
doi: 10.1016/j.bbr.2010.05.014.

Copyright Â© 2010 Elsevier B.V.

Fig. 1. Screenshot of a figure (Figure 2 in Pubmed ID 20570596) in NCBI Images with a summary added by the figure summarizer browser extension. When the extension is not applied, the summary is not visible.

the National Center for Biotechnology Information (NCBI) image database (Canese, 2010), which incorporates 3 million figures (accessed November 6, 2010). In this study, we developed browser extensions to enrich the figure with figure summary.

2 IMPLEMENTATION

We implemented figure summarization program into extensions for the following web browsers: Mozilla Firefox, Google Chrome and Apple Safari (extensions available from: http://figuresearch.askhermes.org/articlesearch/extensions.php). The extensions are written in JavaScript and implemented using Ajax. The summaries have already been generated and are stored on our servers. When the user loads a figure on PubMed Central, the extension makes a call to our server with the article's PubMed Central ID and the figure's ID. If the summary for the figure is available, our server sends it to the extension. The extension then injects HTML code under the caption to display the summary (Figure 1). As the summaries have been preprocessed, no time is spent on generating the summaries. We use Ajax to obtain and display the summary as this allows the figure page to load even if the summary is unavailable or while the summary is obtained from the server. As the extension is based on JavaScript, we have also made it available as a Greasemonkey user script (http://userscripts.org/scripts/show/90065).

The web browser extensions we developed insert a summary for the figure under the caption of the figure (Figure 1). The summary is retrieved from figuresearch's servers by the extensions using Ajax. If an article is not open-access or has not been processed and stored in figuresearch's servers, then the figure is displayed as it would have been without the extension. We believe that use of this extension will allow biomedical researchers to access figure summaries in their traditional workflow, which involves use of PubMed Central.

2.1 Limitations

As only the open-access subset of PubMed Central is freely available for download, figuresearch includes summaries for figures from these open-access articles only. As a result, summaries are displayed for figures from open-access articles only.

Funding: National Institutes of Health [1R21RR024933-01A1 and 5R01LM009836-02] and University of Wisconsin-Milwaukee's Research Growth Initiative, (to H.Y.).

Conflict of Interest: none declared.

REFERENCES

Agarwal,S. and Yu,H. (2009a) Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, **25**, 3174–3180.

Agarwal,S. and Yu,H. (2009b) FigSum: automatically generating structured text summaries for figures in biomedical literature. *AMIA Annu. Symp. Proc.*, **2009**, 6–10.

Canese,K. (2010) PubMed display enhanced with images from the new NCBI images database. *NLM Tech. Bull.*, e14.

Das,S. *et al*. (2003) Membrane-binding and activation mechanism of PTEN. *Proc. Natl Acad. Sci. USA*, **100**, 7491–7496.

Lin,C. (2004) ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.

Yu,H. (2006) Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. *AMIA Annu. Symp. Proc.*, **2006**, 834–838.

Yu,H. *et al*. (2009) Are figure legends sufficient? evaluating the contribution of associated text to biomedical figure comprehension. *J. Biomed. Discov. Collab.*, **4**, 1.