BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 22 2009, pages 3031–3032
doi:10.1093/bioinformatics/btp475

Data and text mining

PubMed-EX: a web browser extension to enhance PubMed search with text mining features

Richard Tzong-Han Tsai[1,*], Hong-Jie Dai[2,3], Po-Ting Lai[1], Chi-Hsin Huang[2]

[1]Department of Computer Science and Engineering, Yuan Ze University, Chung Li, [2]Institute of Information Science, Academia Sinica, Nankang, Taipei and [3]Department of Computer Science, National Tsing-Hua University, HsinChu, Taiwan, R.O.C.

Received on March 12, 2009; revised on July 10, 2009; accepted on July 28, 2009

Advance Access publication August 4, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Summary: PubMed-EX is a browser extension that marks up PubMed search results with additional text-mining information. PubMed-EX's page mark-up, which includes section categorization and gene/disease and relation mark-up, can help researchers to quickly focus on key terms and provide additional information on them. All text processing is performed server-side, freeing up user resources.

Availability: PubMed-EX is freely available at http://bws.iis.sinica.edu.tw/PubMed-EX and http://iisr.cse.yzu.edu.tw:8000/PubMed-EX/.

Contact: thtsai@saturn.yzu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Text-mining technology that automatically extracts key information has been utilized to enhance readability of NCBI PubMed or EMBL-EBI search results (Doms and Schroeder, 2005). Unfortunately, compared with the popularity of PubMed or EMBL-EBI's online search interfaces, none of the above services have been widely adopted by the biomedical research community. Part of the problem may be that all of these advanced services require users to navigate to independent web sites, which have their own search interfaces. If enhanced biomedical information retrieval services could retain the ease of the use of PubMed's search interface while allowing optional advanced functionality, they could possibly win over more converts in the biomedical research community.

A promising model for implementing such an approach is by using a browser extension to modify PubMed's actual search page and create a mashup of PubMed's results with extended annotation, relation mapping, external linking, etc. In this article, we outline the development of such an extension, PubMed-EX, which seamlessly integrates the biomedical text-mining annotation of our text-mining web services with the PubMed search web site to provide users with enhanced views of search results. All text-mining annotation and processing are carried out server-side, eliminating large downloads or updates as well as freeing up users' resources.

*To whom correspondence should be addressed.

2 METHODS

2.1 PubMed-EX extension

The core of the PubMed-EX extension is a JavaScript user script, which works in-browser only on the PubMed domain. It updates a user's PubMed search return pages with content processed by our text-mining server. To modify the PubMed pages, the script uses the browser document object model (DOM), which defines the content, structure and style of an HTML document. For example, using the DOM interface, our user script can determine which DOM element contains the abstract text returned by PubMed. It can then extract the raw text from the element, and send it to our text-mining server. Upon receiving the processed results, the user script modifies the appropriate DOM element attribute and the content displayed on the user's browser is updated immediately.

In order to run our script, we use the GreaseMonkey extension for Firefox and the IE7Pro extension for IE. Both are capable of applying custom scripts to pages and can save the effort of writing our own extension code from scratch. To streamline installation, we bundle our script with GreaseMonkey/IE7Pro. End users only need to download and install the appropriate PubMed-EX extension for their browser.

2.2 Text-mining server

The text-mining server comprises three text mining web services and a script proxy, which enables communication between PubMed-EX and the text-mining server.

2.2.1 Section categorizer The section categorizer divides abstracts into section paragraphs. For a given abstract, if the pre-sectioned check finds that the abstract contains obvious section tags, such as 'Objective' and 'Conclusion', the abstract is immediately divided into paragraphs. The pre-sectioned check uses a list of tag keywords to determine whether the abstract is predivided. If the check cannot find any obvious tags, a conditional random fields (CRF)-based categorizer extended from Lin *et al*. (2008) is employed to section the given abstract. Our categorizer achieves a satisfactory F-score of 98.82%.

2.2.2 Named entity taggers The text-mining web services include two named entity (NE) taggers. The first is a machine learning-based NE tagger (Dai *et al*., 2007; Tsai *et al*., 2006), which labels gene names in abstracts. Following NE recognition, a normalization module maps the found gene names to their corresponding Entrez Gene database identifiers using heuristic normalization rules (Lai *et al*., 2009). After the normalization process is completed, successfully normalized gene names are collected into a list, and the abstract is then rechecked for NEs on the list in case the NE tagger missed any instances of the found gene names. Our gene name

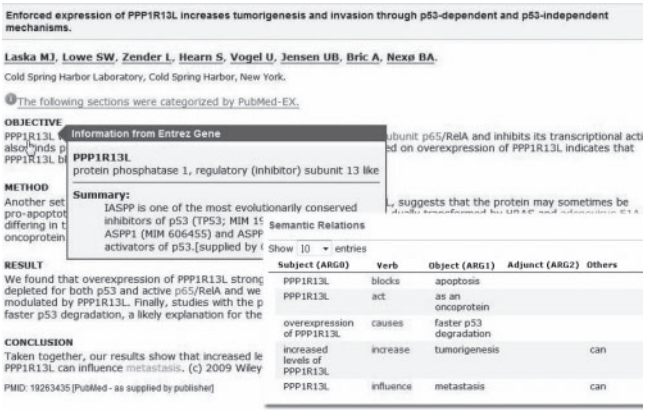

Fig. 1. Query results from PubMed search marked up by PubMed-EX.

tagger and normalization system achieve *F*-scores of 86.24% and 83.4% on BioCreAtIvE II.

The second NE tagger is a dictionary-based disease name tagger based on the maximum matching algorithm. We compiled a dictionary of about 40 000 disease terms with corresponding unique identifiers from the MeSH database. It achieves satisfactory *F*-score of 83.4% on Jimeno *et al*.'s (2008) corpus. Detailed results can be found in the Supplementary Material.

2.2.3 Relation analyzer The relation analyzer extracts relations among phrases and biomedical verbs in sentences using the semantic role labeling (SRL). In SRL, sentences are represented by one or more predicate–argument structures (PAS; Wattarujeekrit *et al*., 2004). Each PAS is composed of a predicate (e.g. a verb) and several arguments that have different semantic roles, including main arguments such as agent (subject) or patient (object), as well as adjunct arguments, such as time or location. Take the phrase 'Because the variant of the NeuroD/BETA2 gene is associated with type 1 but not type 2 diabetes', for example, 'NeuroD/BETA2 gene' and 'with type 1 but not type 2 diabetes' are identified as object and adjunct arguments in the PAS whose verb is 'associate'. Our relation analyzer, which is based on our previous SRL system BIOSMILE (Tsai *et al*., 2007), achieves an *F*-score of 81.75%.

2.2.4 Script proxy PubMed-EX's user script is written in JavaScript. However, to avoid malicious scripts, hacks and exploits, most browser script engines block scripts from directly accessing off-site web services outside the domain of the current page. As a result, PubMed-EX cannot directly communicate with our text-mining web services because they are outside the PubMed domain. The script proxy server offers PubMed-EX, a pathway around this restriction. The idea is simple: PubMed-EX sends its service requests in the form of HTTP requests (which are not restricted by browser script engines) to the proxy server. The proxy server then translates the HTTP requests into SOAP remote procedure calls, forwards them to our text-mining web services and receives, retranslates and returns the results via HTTP to PubMed-EX.

3 USAGE AND INSTALLATION

Figure 1 shows the search results marked up by PubMed-EX. As you can see, the uncategorized abstract is automatically sectioned by PubMed-EX into objectives, methods, results and conclusions. Automatically sectioned abstracts are preceded by a note to inform researchers that they were categorized by PubMed-EX. Biomedical NEs in the displayed results are marked in different colors. When the users move their mouse cursors over recognized NEs (e.g. PPP1R13L), a brief pop-up summary will be displayed. The recognized genes (or diseases) also hyperlink to corresponding Entrez Gene (MeSH) pages.

After installing PubMed-EX and choosing the 'AbstractPlus' view from PubMed's pull-down display menu, users can see semantic relation analysis in the lower-right pane. As you can see in Figure 1, the semantic relation analysis results are shown under 'Related Articles' in table format. (Users can click on the header, 'Semantic Relations', to expand the lower-right pane to cover the entire right-hand side.)

All sentences that contain more than one NE are analyzed by our relation analyzer and marked up by PubMed-EX. For example, action verbs are highlighted in pink. When users move their mouse cursor over a relational element, the corresponding sentence in the abstract on the left-hand side is highlighted in green. The major relational elements, including subject, negation, verb, adjunct, object, location, time and extent are listed in different columns in the right-hand pane.

Users can visit http://bws.iis.sinica.edu.tw/PubMed-EX or http://iisr.cse.yzu.edu.tw:8000/PubMed-EX/ to download the extension.

Conflict of Interest: none declared.

REFERENCES

Dai,H.-J. *et al.* (2007) IASL systems in the gene mention tagging task and protein interaction article sub-task. In *Proceedings of Second BioCreAtIvE Challenge Evaluation Workshop*. CNIO (Spanish National Cancer Research Centre), Madrid, Spain, pp. 69–76.

Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33**, W783–W786.

Jimeno,A. *et al.* (2008) Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, **9**, S3.

Lai,P.-T. *et al.* (2009) Using contextual information to clarify gene normalization ambiguity. In *The IEEE International Conference on Information Reuse and Integration (IEEE IRI 2009)*. Las Vegas, USA.

Lin,R.T.K. *et al.* (2008) Result identification for biomedical abstracts using conditional random fields. In *The IEEE International Conference on Information Reuse and Integration (IEEE IRI 2008)*. Las Vegas, Nevada, USA.

Tsai,R.T.-H. *et al.* (2006) NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, **7**, S11.

Tsai,R.T.-H. *et al.* (2007) BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, **8**, 325.

Wattarujeekrit,T. *et al.* (2004) PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, **5**, 155.