Proxiris, an augmented browsing tool for literature curation

Vahé Chahinian, Marie-Jean Meurs, David H. Mason, Erin McDonnell, Ingo Morgenstern, Greg Butler, Adrian Tsang

Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

Abstract. We present proxiris, an augmented browsing tool supporting the manual curation of literature related to genomics-based lignocellulose research. In web pages, proxiris highlights information automatically retrieved by the mycoMINE text mining system. It provides researchers and biocurators with an overview of the document content along with additional information and links to external resources. The proxiris development and implementation are based on a proxy server approach, allowing a seamless integration with the user's browser.

1 Introduction

Electronic scientific publications available in multiple repositories reach an overwhelming amount and continue to grow steadily. Regarding PubMed [9], the largest knowledge source available to biological researchers, more than 22 million articles were indexed as of March 2013. Accessing this information is crucial for conducting research and designing experiments. However, querying various biological bibliographic databases for retrieving data of particular interest often produces a long list of potentially relevant papers. Reading these papers to extract critical information is an unavoidable step in the literature curation process which supports research. Unfortunately, this step is a bottleneck in the knowledge discovery workflow [5] since the task is highly time-consuming and error-prone.

Several tools have been developed to help researchers and literature curators. Among them are Reflect, PubMed-Ex and PubTator. Reflect[1] [8] is a free service that can be installed as a plug-in in Mozilla Firefox, Internet Explorer, or Google Chrome web-browsers. Reflect tags only gene, protein and small molecule names in web pages. Clicking on a tagged term opens a pop-up showing summary information on the term and related Wikipedia content if available. PubMed-EX[2] [11] is a browser extension that marks up PubMed bibliographic database [9] search results with additional text-mining information. PubMed-EX page mark-up includes section categorization, gene/disease name, and relation. PubTator [12] is a web-based tool that allows curators to create, save, and export annotations. Using the Entrez API, PubTator allows the same search syntax and returns

[1] `http://reflect.ws`
[2] `https://sites.google.com/site/hongjiedai/projects/pubmed-ex/`

identical search results as PubMed. It highlights genes, chemicals, diseases, and species in the titles of retrieved papers and also in their abstracts using state-of-the-art text mining tools, e.g. SR4GN[13] for species recognition and gene normalization.

All the aforementioned tools significantly improve the user experience when browsing the web. However, none of them provides users with seamlessly integrated yet flexible natural language processing services. Our motivations for building a new system are based on the need of user friendly systems for biocurators [4] which (i) provide additional information while preserving original content and format of browsed web pages, (ii) allow user interaction, (iii) can be easily adapted to various research fields. Our system, proxiris, is a web-based tool developed to support users who need to mine huge volumes of web publications. Proxiris is an open source project designed to meet all three requirements but only the first one is currently implemented. The next section describes services provided by proxiris to its users. Section 3 presents proxiris architecture and implementation, then Section 4 discusses results and future improvements.

2 Description

Proxiris is a web-based tool developed at the Centre for Structural and Functional Genomics for supporting researchers, curators and experimenters working towards discovery and development of effective fungal enzyme cocktails which can convert lignocellulose into fermentable sugars. The manual curation of fungal genes encoding lignocellulose-active enzymes is an essential step for supporting this research, as it allows researchers to easily access reliable knowledge. When browsing through literature, researchers mainly read fragments of interesting papers (for instance, sections describing methods or results). Biocurators study relevant papers with an exhaustive approach since their goal is translation and integration of relevant information into a database. As shown in Figure 1,

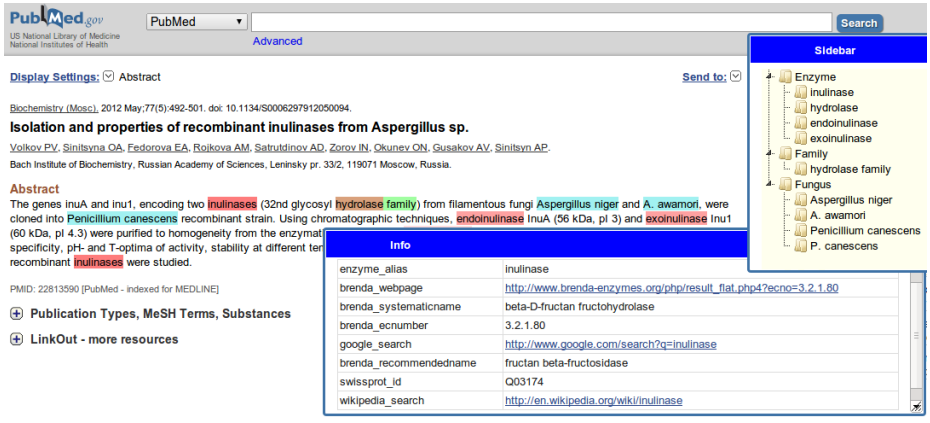

Fig. 1. Proxiris running in Google Chrome

proxiris eases literature mining for both categories of readers by highlighting entities of interest that are listed in an interactive sidebar. For these entities,

it also gives access to added content from external databases in the form of identifiers or direct links to web pages. Proxiris augmented browsing relies on the mycoMINE [6] text mining system. Retrieved fungal names are mapped to their NCBI Taxonomy identifiers and web pages. Retrieved enzyme names are mapped to their BRENDA [2] identifiers, recommended and systematic names, and their UniProt [10] identifiers. They are also classified according to the enzyme family they belong to using CAZy [1]. mycoCLAP [7], our on-site developed database of fungal genes encoding lignocellulose-active proteins, is used as a complementary knowledge source for recognition of assay, fungus, gene, and substrate names. mycoMINE is implemented using the GATE [3] framework. Domain vocabularies are based on gazetteer lists and ontologies. Links to external resources can be manually defined in dedicated processing resources implemented for suiting mycoMINE user needs. For supporting various research areas, mycoMINE can be replaced by other text mining systems that process content in HTML format.

3 Implementation

Proxiris implementation is based on a proxy server approach depicted in Figure 2. The user's **browser** is set to use our **proxiris proxy** server that acts as an intermediary for http requests from users seeking literature web pages from other servers (**Origin** in Fig.2, e.g. PubMed). The user's http request goes to the proxy server which then retrieves the requested page. The proxy is a node.js service that processes the page by sending appropriate parts of the HTML to the mycoMINE pipeline [6] and then replacing the original content with the processed text. The proxiris **instrumentation** is implemented in JavaScript using node.js,

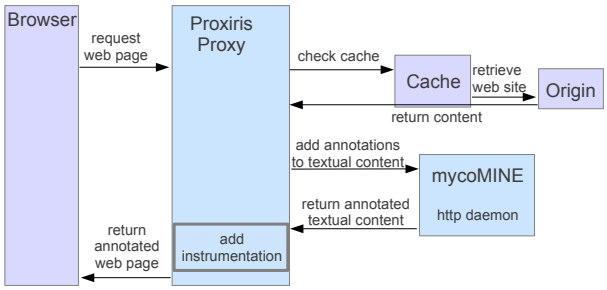

Fig. 2. Proxiris architecture

jsTree and CSS. node.js is a lightweight event driven server-side JavaScript system. As JavaScript is already being used for the front end, almost all the programming is done in one language. This uniform approach makes it easier to shift processing between the server and client ("node" refers to a distributed network deemphasizing the client-server model). In proxiris, node.js processes the HTML document returned by the text mining pipeline to support the highlighting functionality. Different document sections or even processes can be mapped to different web addresses. node.js is also used to inject the proxiris JavaScript and CSS into the original webpage. JavaScript is used to implement all of the dynamic features such as the left click and the highlight toggle sidebar functionality. jsTree

is a library which builds the sidebar in a very simple way. CSS is used to provide all the HTML markup descriptions. CSS is used for highlight colours and the dialog box colours and shape for proxiris. The highlight colours are implemented as different CSS classes.

4 Conclusion

Two curators evaluated a prototype of proxiris on the triage of 114 PubMed abstracts. Using the tool, the time needed for triage was reduced by 21%, showing the relevance of the approach. The proxy approach supporting proxiris is flexible and generic enough to easily integrate various text mining and semantic services. A proxy also obviates the need to limit browser selection or to write custom browser code for all popular browsers. Not only does this approach preserve the format of the original document (including pictures, tables and embedded services), it also circumvents the same origin policy which keeps browser scripts loaded from one origin from interacting with a resource from another origin. Using a proxy enables caching of publications from selected websites which can be quickly retrieved by the user. This approach also provides reliable services for users since proxiris is enabled only on selected web sites with correctly handled content. Our current work is focused on user interaction with the application to document triage and curation.

Acknowledgments.
Funding for this work was provided by Genome Canada and Génome Québec.

References

1. Cantarel et al.: The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic acids research 37(suppl 1) (2009)
2. Chang et al.: BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Research 37(suppl 1) (2009)
3. Cunningham et al.: Text Processing with GATE (Version 6). University of Sheffield, Dept. of Computer Science (2011)
4. Hirschman et al.: Text mining for the biocuration workflow. Database 2012 (2012)
5. Howe et al.: Big data: The future of biocuration. Nature 455, 47–50 (2008)
6. Meurs et al.: Semantic text mining support for lignocellulose research. BMC Medical Informatics and Decision Making, Vol 12 Suppl 1 (2012)
7. Murphy et al.: Curation of characterized glycoside hydrolases of fungal origin. Database: The Journal of Biological Databases and Curation 2011 (2011)
8. Pafilis et al.: Reflect: augmented browsing for the life scientist. Nature Biotechnology 27, 508–510 (2009)
9. Sayers et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 38(suppl 1), D5–D16 (2009)
10. The UniProt Consortium: The Universal Protein Resource (UniProt). Nucleic Acids Research 37(D), 169–174 (2009)
11. Tsai et al.: PubMed-EX: a web browser extension to enhance PubMed search with text mining features. Bioinformatics 25(22), 3031–3032 (2009)
12. Wei et al.: Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in pubmed abstracts. Database 2012 (2012)
13. Wei et al.: SR4GN: a species recognition software tool for gene normalization. PloS one 7(6) (2012)