BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 23 2011, pages 3327–3328
doi:10.1093/bioinformatics/btr564

Data and text mining

Advance Access publication October 12, 2011

BioTextQuest: a web-based biomedical text mining suite for concept discovery

Nikolas Papanikolaou[1,2,†], Evangelos Pafilis[2,†], Stavros Nikolaou[1], Christos A. Ouzounis[3,‡], Ioannis Iliopoulos[2,*] and Vasilis J. Promponas[1,*]

[1]Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, CY 1678, Nicosia, Cyprus, [2]Division of Medical Sciences, University of Crete Medical School, Heraklion 71110, Greece and [3]Centre for Bioinformatics, Department of Computer Science, School of Physical Sciences and Engineering, King's College London, London, UK

Associate Editor: Alex Bateman

ABSTRACT

Summary: BioTextQuest combines automated discovery of significant terms in article clusters with structured knowledge annotation, via Named Entity Recognition services, offering interactive user-friendly visualization. A tag-cloud-based illustration of terms labeling each document cluster are semantically annotated according to the biological entity, and a list of document titles enable users to simultaneously compare terms and documents of each cluster, facilitating concept association and hypothesis generation. BioTextQuest allows customization of analysis parameters, e.g. clustering/stemming algorithms, exclusion of documents/significant terms, to better match the biological question addressed.

Availability: http://biotextquest.biol.ucy.ac.cy

Contact: vprobon@ucy.ac.cy; iliopj@med.uoc.gr

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

Received on May 23, 2011; revised on September 15, 2011; accepted on October 6, 2011

1 INTRODUCTION

Several information retrieval tools are available to the scientific community, each attempting to analyze the large amount of biomedical information from a different perspective (Kim and Rebholz-Schuhmann, 2008). Most of the currently available literature mining systems are based on Natural Language Processing, supervised machine learning or their combination (Cohen and Hunter, 2008; Larranaga *et al.*, 2006; Winnenburg *et al.*, 2008). An alternative approach is the summarizing of large query outputs by document clustering, where similar documents are grouped on the basis of particular, meaningful terms (Iliopoulos *et al.*, 2001). It is an unsupervised technique, thus its success does not depend on prior knowledge or domain expertise (Iliopoulos *et al.*, 2001; Lu, 2011; Theodosiou *et al.*, 2008). Existing biomedical literature clustering systems include ClusterMed™ (http://demos.vivisimo.com/clustermed) and PuReD-MCL (Theodosiou *et al.*, 2008). ClusterMed™ is a commercial software whose free version supports the clustering of at most 100 abstracts. PuReD-MCL is a downloadable collection of Perl and R scripts, mostly addressing expert users (Theodosiou *et al.*, 2008). With BioTextQuest, our aim has been to provide researchers with a user-friendly web application that enables online queries for defining corpora of biomedical abstracts followed by document clustering and visualization. BioTextQuest is an application that offers an array of visualization tools for efficient navigation among biomedical records, and concept extraction. It also supports the depiction of term associations within clusters and can be easily customized. In the following sections, we describe the BioTextQuest architecture and demonstrate its functionality, based on a reworked query example of a *Drosophila* developmental pattern presented in the original TextQuest publication (Iliopoulos *et al.*, 2001).

2 IMPLEMENTATION

BioTextQuest is a user-friendly web application, offering a simple interface and support for visual term associations. The core component is based on TextQuest (Iliopoulos *et al.*, 2001), a prototype capable of extracting the terms of biomedical significance in a group of records and grouping these records according to their similarity based on the extracted terms. To query PubMed, while minimizing response times and hardware resource requirements, BioTextQuest (i) uses the NCBI eSearch service from the Entrez Programming Utilities (http://eutils.ncbi.nlm.nih.gov/) to retrieve the identifiers of the PubMed records matching a given user query and (ii) looks up these identifiers in a local index that matches them to their corresponding textual contents. Such a system takes advantage of the PubMed query syntax and of the capabilities of the NCBI utilities while mitigating the necessity of a local full-text PubMed index. A daemon monitors the new MEDLINE/PubMed releases and is responsible for downloading the new records and dynamically updating the local index. An overview of the BioTextQuest architecture is provided in Supplementary Figure S1, and it has been tested on the most common web browsers (Firefox, Chrome, Safari and Opera). Compared with the original TextQuest algorithm, the web application BioTextQuest:

(1) offers optional use of the Lovins stemming algorithm (Lovins, 1968);

(2) employs a set of clustering algorithms namely: (i) *K*-Means [Cluster 3.0 package, de Hoon *et al.* (2004)], (ii) spectral clustering

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

‡Present addresses: Computational Genomics Unit, Institute of Agrobiotechnology, Center for Research & Technology Hellas (CERTH), GR-57001 Thessaloniki, Greece; Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M5S 3E1, Canada.

and hierarchical average linkage clustering, [`clusterx` command line interface to SCPS, Nepusz *et al.* (2010)] and (iii) Markov Clustering - MCL, a graph clustering algorithm (Enright *et al.*, 2002);

(3) offers integrated visualization and interactive re-analysis capabilities, such as the option of selecting terms, documents and clusters for removal and re-clustering. User interface elements are based on HTML, CSS and JavaScript (JavaScript libraries: scriptaculous (http://script.aculo.us/), overlibmws (http://www.macridesweb.com/oltest/);

(4) invokes third-party text annotation web services [Reflect, Pafilis *et al.* (2009); WhatIzIt, Rebholz-Schuhmann *et al.* (2008)] to annotate the terms of biomedical importance identified by the clustering algorithm.

3 FUNCTIONALITY

The BioTextQuest server is designed to be as user friendly as possible, offering the aforementioned features in the most intuitive way. A Google-like search page allows users to pose direct queries to PubMed and interactively set the analysis parameters. Text annotation services are employed to map the extracted terms of biomedical significance to the biological entities they describe. The resulting document clusters and their corresponding terms, along with the retrieved annotation, are combined to produce a series of views of the analysis results. Importantly, users can interactively subcluster and/or recluster the results of any completed analysis.

The Home page (Supplementary Fig. S2) is designed to be minimal to allow users to concentrate on the main functionality. Pop-up instructions assist users in executing an analysis. More complex queries can be issued following the Entrez syntax. Users may specify up to a maximum of 2000 articles to be retrieved/processed. Advanced options, such as stemming and clustering methods (with respective parameters), are hidden by default and may be used to fine-tune the results. Each algorithm may prove useful to address different kinds of questions. Since the functionality of the web tool partly relies on other web services (e.g. Entrez Programming Utilities), users are notified whether these services might be unreachable.

The Results page (Supplementary Figs S3–S5) provides views organized under tabs, along with a frame displaying a summary (query and parameters) and a form for a new run. The 'Tag Clouds' view returns an informative display of cluster-specific terms (Supplementary Fig. S3). This type of literature mining for biomedical concept discovery (Krallinger *et al.*, 2008), reflects the size of each term proportional to the fraction of documents with the specific term within the cluster. The 'Biomedical Terms' view lists and graphically annotates the GoList terms identified in the same manner as above (Supplementary Fig. S4). Finally, the 'Documents' view provides titles of documents belonging to each cluster with a PubMed link, and links to the plain or semantically annotated (Reflected) abstract (Supplementary Fig. S5). Users may interact with the results by requesting a fine-grained analysis of the active corpus, i.e. the collection of abstracts currently being processed. This is achieved by any combination of (i) simply altering available options (e.g. stemming or clustering); (ii) excluding terms from the GoList available in the Biomedical Terms view and (iii) removing clusters of documents from the Tag Clouds view. BioTextQuest employs third-party text annotation web services to enrich terms (by highlighting terms and using popups) according to the biological entity they describe. To identify protein names BioTextQuest invokes the Reflect service (Pafilis *et al.*, 2009). WhatIzIt (Rebholz-Schuhmann *et al.*, 2008) is optionally used to identify terms referring to pathway, molecular function and/or cellular component names. Additionally, the user can highlight non-standard English terms and terms unique to a cluster. When stemming is used, the root terms are displayed in the Tag-Cloud, whereas the respective unstemmed terms are available upon clicking on the respective link.

4 CONCLUSION

We describe BioTextQuest, an online text mining system, enabling concept discovery from the biomedical literature recorded in PubMed via a clustering-based procedure. Further exploration of literature information is facilitated by intuitive graphical depictions of biologically relevant terms and interactive re-/subclustering capabilities. BioTextQuest is designed with simplicity for end-users, while advanced options enable further analyses.

ACKNOWLEDGEMENTS

Many thanks to BRL (University of Cyprus) and CBG (University of Crete) members for testing and feedback, and the three anonymous reviewers for valuable comments.

Funding: Cyprus Research Promotion Foundation [Grant YGEIA/BIOS/0308(BE)/11]; MICROME, a Collaborative Project funded by the European Commission within its FP7 Programme, Contract No. (222886-2); MARBIGEN EU FP7 REGPOT Project (Reference: 264089 to E.P.), in part; INFLACARE EU FP7 Health Project (Reference: 223151 to N.P.), in part.

Conflict of interest: none declared.

REFERENCES

Cohen,K.B. and Hunter,L. (2008) Getting started in text mining. *PLoS Comput. Biol.*, **4**, e20.

de Hoon,M.J. *et al.* (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Iliopoulos,I. *et al.* (2001) Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput.*, 384–395.

Kim,J.J. and Rebholz-Schuhmann,D. (2008) Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Brief. Bioinformatics*, **9**, 452–465.

Krallinger,M. *et al.* (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Larranaga,P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinformatics*, **7**, 86–112.

Lovins,J.B. (1968) Development of a stemming algorithm. *Mechanical Translation and Comput. Ling.*, **11**, 22–31.

Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*; [Epub ahead of print, doi:10.1093/database/baq036, January 18, 2011].

Nepusz,T. *et al.* (2010). SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC Bioinformatics*, **11**, 120.

Pafilis,E. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.

Rebholz-Schuhmann,D. *et al.* (2008) Text processing through Web services: calling Whatizit. *Bioinformatics*, **24**, 296–298.

Theodosiou,T. *et al.* (2008) PuReD-MCL: a graph-based PubMed document clustering methodology. *Bioinformatics*, **24**, 1935–1941.

Winnenburg,R. *et al.* (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief. Bioinformatics*, **9**, 466–478.