wwwTextQuest: A Biomedical Text Mining Suite for Concept Discovery

Nikolaos Papanikolaou[1,2+], Evangelos Pafilis[2+], Stavros Nikolaou[1], Ioannis Iliopoulos[2*], Vasilis Promponas[1*]

[1] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, P.O. Box 20537, CY 1678, Nicosia, Cyprus

[2] Department of Basic Research, Medical School, University of Crete, Heraklio, Greece.

[+]: these authors contributed equally to this work

*Correspondence to: vprobon@ucy.ac.cy, iliopj@med.uoc.gr

Motivation: The overwhelming accumulation of biomedical information raises great challenges for information retrieval. Querying literature databases such as Medline[1] only allows a limited exploitation of the knowledge available[2]. User queries often result in long and non-useful lists of matching records. TextQuest [2] is a prototype of a large-scale document clustering system that offers the grouping of biological text records, extracting terms of biomedical significance. Based on these features, a long list of results is replaced by groups of related records and by a graph of terms that co-exist-in and characterize each group. Aim of this project is to make TextQuest available as a web application, allowing (i) interactive navigation among clusters of biomedical records relevant to a user defined query, (ii) support graphical associations among the terms of the clusters, and (iii) provide an intuitive and user friendly graphical user interface.

Methods: wwwTextQuest collects abstracts from the Medline literature database, or records from the OMIM human genetic disorders database, matching a user query. It identifies bibliometrically relevant terms, calculates pairwise document similarities (using the Vector Space Model). Employing an array of clustering algorithms, it transforms results into clusters of records and of their corresponding terms. The web application follows a three-tier architecture (Fig. 1) that besides the document processing and clustering algorithms, relies on public web services such as NCBI eSearch, Reflect [3] and WhatIzIt [4] to query biomedical databases and to annotate and enrich the biomedically significant terms.

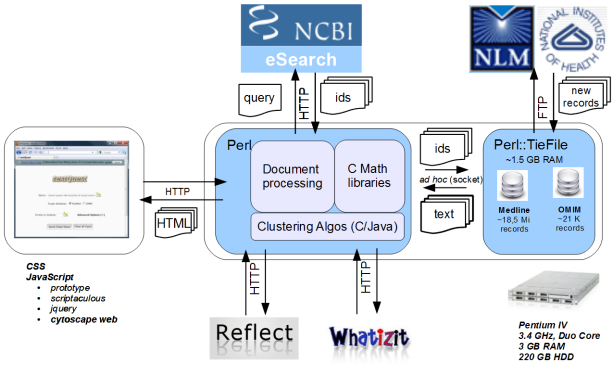

Figure 1: wwwTextQuest architecture.

Results: TextQuest was converted from a standalone prototype to a user-friendly web service. Not only does it employ a greater variety of clustering and stemming algorithms, but it also achieves interactive result navigation. In wwwTextQuest, users start with a query as they normally would in Medline or OMIM. They can define the clustering and stemming algorithm to be used along with their parameters (Fig. 2a). Results are presented in four different views, namely: document clusters, tag clouds of the terms of each cluster (Fig 2b), interactive term graphs and the list of the extracted terms of biomedical significance ("GoList"). To enhance these views even further, the extracted terms are annotated according to the biological entity they describe (gene, protein, pathway, molecular function, and/or cellular components). Finally, as a further improvement, the whole functionality is offered recursively, i.e. by simply using the web interface, users can sub-cluster and/or re-cluster, clusters from previous analyses.

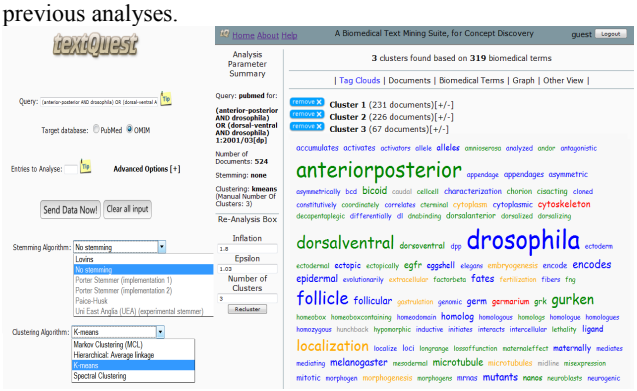

Figure 2: (a) Query Page, (b) Result Page

Discussion: Document clustering is capable of grouping biomedical records and extract biomedical concepts [1]. wwwTextQuest is a web-based biomedical literature mining suite able to satisfy the needs for efficient navigation among biomedical records for concept extraction and association.

State-of-the-art visualization methods allow to simultaneously compare the terms characterizing the document clusters. Such a comparison may serve as the basis for hypothesis formulation (e.g. a gene to a developmental stage association) and paves the way towards knowledge discovery.

To this purpose users are assisted by the graphical annotation of different biological entity types and, even more, by the ability to explore iteratively subsections of previous analyses by just using the graphical interface.

Current work is focusing on fine-tuning the analysis parameters and evaluating the performance of wwwTextQuest (still an alpha version) on real life problems. In the future we intend to extend the system by incorporating more clustering and stemming algorithms, as well as data from more resources.

Acknowledgements: This work is co-funded by the Republic of Cyprus and the EU European Regional Development Fund (ERDF) through a grand [ΥΓΕΙΑ/ΒΙΟΣ/0308(BE)/11] from the Cyprus Research Promotion Foundation to the authors.

References

1. Iliopoulos I *et al,*., **TextQuest: Document Clustering of Medline Abstracts for Concept Discovery in Molecular Biology.** *Pac Symp Biocomp*, 2001, 384-395
2. Medline: http://www.nlm.nih.gov
3. Reflect: http://reflect.cbs.dtu.dk
4. WhatIzIt: http://www.ebi.ac.uk/webservices/whatizit/