BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access

www.bioinformation.net

Software

Volume 8(12)

TPX: Biomedical literature search made easy

Thomas Joseph, Vangala G Saipradeep, Ganesh Sekar Venkat Raghavan, Rajgopal Srinivasan*, Aditya Rao, Sujatha Kotte & Naveen Sivadasan#

TCS Innovation Labs - Hyderabad, Tata Consultancy Services, 1, Software Units Layout, Madhapur, Hyderabad - 500081, INDIA; # Current affiliation: Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Ordnance Factory Estate, Yeddumailaram 502205, India; Rajgopal Srinivasan – Email: raj@atc.tcs.com; Phone: +91-40-6667 3551; *Corresponding author

Received May 29, 2012; Accepted June 08, 2012; Published June 28, 2012

Abstract:
TPX is a web-based PubMed search enhancement tool that enables faster article searching using analysis and exploration features. These features include identification of relevant biomedical concepts from search results with linkouts to source databases, concept based article categorization, concept assisted search and filtering, query refinement. A distinguishing feature here is the ability to add user-defined concept names and/or concept types for named entity recognition. The tool allows contextual exploration of knowledge sources by providing concept association maps derived from the MEDLINE repository. It also has a full-text search mode that can be configured on request to access local text repositories, incorporating entity co-occurrence search at sentence/paragraph levels. Local text files can also be analyzed on-the-fly.

Availability: http://tpx.atc.tcs.com/

Keywords: PubMed search, text-mining, concept identification, biomedical literature, concept association, concept-assisted search, ontology based dictionaries.

Background:
PubMed, the most popular and publicly available life science literature retrieval tool, is generally used for retrieval of specific information from MEDLINE rather than as an exploratory medium. There are several other retrieval tools for searching MEDLINE such as GoPubMed **[1]** and EBIMed **[2]** that have been designed to provide improved retrieval from literature and other related information sources. However, it is desirable to have a mechanism which (a) utilizes the strengths of PubMed (b) enables users to search, explore and manage the literature more effectively and (c) enables integration with structured knowledge sources. We have developed TPX (**T**CS **P**ubmed e**X**plorer), a web-based tool which supports concept-assisted search and navigation that relies on PubMed as the underlying search engine to search the MEDLINE database.

In addition, many users have local collections of free/purchased articles as well as other text documents related to their specific areas of research in the biomedical domain. Exploring and information retrieval from these collections for purpose of extracting specific information remains a challenge. TPX can be configured on request to give the user the ability to explore the user's own locally available collection of articles. To demonstrate this option, TPX has been configured to access the Open Access subset of PMC articles using the full-text search mode. In addition, while searching through local repositories,

users can limit co-occurring terms to be within the same sentence, paragraph or the entire article.

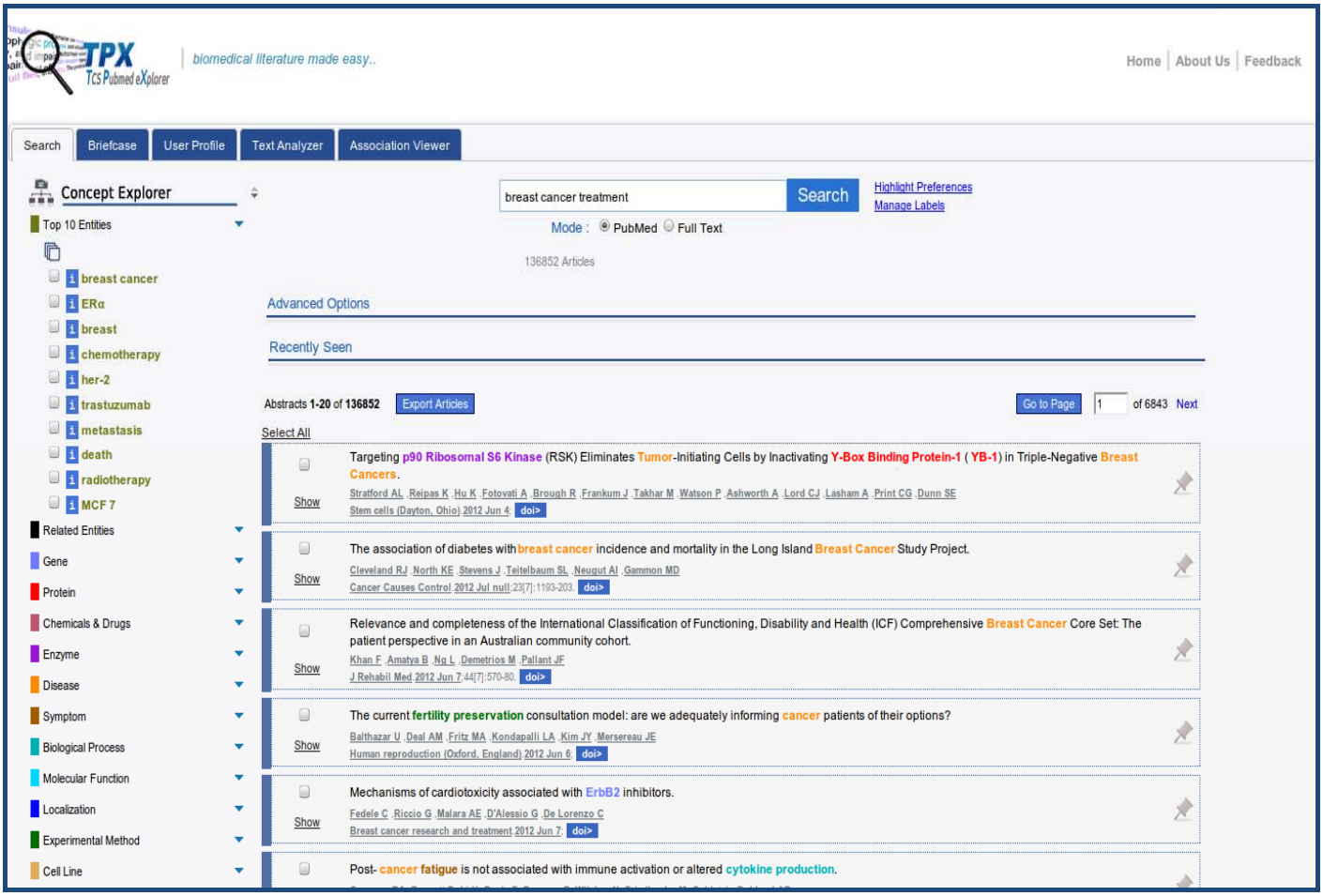

Figure 1: Shows a screen-shot of the tool

Tool features:

Concept Identification

TPX is equipped with a dictionary-based Named Entity Recognition (NER) system that identifies various biological concepts in an article using approximate string matching rules. The NER system identifies a wide range of concept categories such as genes, proteins, diseases, symptoms, chemicals and drugs, processes, functions, localization, experimental methods and cell lines derived from ontologies like MeSH **[3],** Gene Ontology **[4]** and other sources like Entrez Gene **[5]**, UniProt **[6]**, Expasy Enzyme **[7]**, NCBI Taxonomy **[8]** and HyperCLDB **[9]**. The dictionaries are modular - so that specified types can be selectively used and dictionaries can be added/removed on request. Local abbreviation-handling is also integral to the tool **[10].** Individual users can further include custom categories and concepts to augment the provided dictionaries. These custom concepts are recognized in articles and also used for identifying co-occurrences in the full-text mode. SNPs and other mutation mentions are currently tagged and highlighted on-the-fly using a pattern-based NER system.

Tagged Abstracts

The abstracts retrieved in PubMed/full-text mode are displayed with the identified entities highlighted with predefined colors **(Figure 1).**

The properties for each of the identified entities can be viewed by clicking it. The properties include external links to respective information sources, facility to explore associated concepts, feedback, etc. The tool provides a facility for the users to suggest entities that have been missed out by the NER system. Through user preferences, the user can customize highlighting only those concepts belonging to categories of interest.

Concept Assisted Search and Navigation

The identified concepts from the analyzed article set of the search results are ranked on-the-fly according to their relevance to the article collection. The ranked concepts are categorized under the respective categories allowing the user to explore the articles with respect to the categorized concepts. These concepts can be exported and saved into file in various formats. Users can select concepts of interest and this selection can either be used to filter the result or be used in query refinement. Entity co-occurrences (at sentence or paragraph level) in full-text articles can be searched in the full-text mode, using terms in the system dictionaries as well as user-created terms.

Concept Association Map

The tool has a pairwise concept association map incorporated. These associations are pre-computed and ranked according to their relevance to the whole of the tagged MEDLINE corpus.

The user can explore the associations starting from a concept of interest by selecting it in the tagged abstract. For each association, a set of relevant article abstracts are displayed. In addition, through the Concept Panel or Association Viewer, users can retrieve sub-maps of the association map such as associations between proteins and chemicals related to a disease like breast cancer. These sub-maps can be exported and saved into file as a list of associations.

Notes, Comments, Labels and Pinned Abstracts

Users can store personal notes and comments for each abstract. Notes are specific to a user, while comments are visible to all users. Users can bookmark articles of interest using custom labels. Alternatively, the 'article pinning' option allows users to bookmark important articles without attaching specific labels to them. All such data can be viewed under the user's Briefcase.

Conclusion:

We have developed TPX, a search analysis and exploration tool whose various features aid in faster information hunt in addition to enhanced exploration and management of PubMed/local repository search results. The user terms/types feature makes TPX an ideal "personal information assistant" for scientists and biomedical literature curators. We believe that relying on PubMed as the basic search/indexing engine and having an advanced search analysis tool that performs customized off-line and on-the-fly analysis of the search result is an effective way of integrating biomedical literature with in-house and other external knowledge sources. TPX, with the ability to point to external websites in addition to local repositories is a step towards faster, personalized biomedical literature analysis.

References:

[1] Doms A & Schroeder M, *Nucleic Acids Res*. 2005 **33**: W783 [PMID: 15980585]

[2] Rebholz-Schuhmann D *et al. Bioinformatics*. 2007 **23**: e237 [PMID: 17237098].

[3] http://www.nlm.nih.gov/mesh

[4] Ashburner M *et al. Nat Genet*. 2000 **25**: 25 [PMID: 10802651]

[5] Maglott D *et al. Nucleic Acids Res*. 2011 **39**: D52 [PMID: 21115458]

[6] The UniProt Consortium, *Nucleic Acids Res*. 2011 **39**: D214 [PMID: 21051339]

[7] Bairoch A, *Nucleic Acids Res*. 2000 **28**: 304 [PMID: 10592255]

[8] Sayers EW *et al. Nucleic Acids Res*. 2011 **39**: D38 [PMID: 21097890]

[9] Romano P *et al. Nucleic Acids Res*. 2009 **37**: D925 [PMID: 18927105]

[10] Ao H & Takagi T, *J Am Med Inform Assoc*. 2005 **12:** 576 [PMID: 15905486]

Edited by P Kangueane

Citation: **Joseph *et al.*** Bioinformation 8(12): 578-580 (2012)