Visualizing Semantic Metadata from Biological Publications

Johannes Hellrich, Erik Faessler, Ekaterina Buyko and Udo Hahn

Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
http://www.julielab.de

Abstract. The biomedical domain has several large-sized, conveniently searchable document stores, such as MEDLINE or PMC. These collections count up to several millions of publications, with manually supplied, value-adding index terms. Increasingly, these repositories undergo computational text analysis which yields additional, automatically generated semantic metadata. Due to the volume and complexity of these data and their interrelations, standard textual modes for displaying search results become more and more inappropriate. Therefore, we propose an interactive graphical front-end, embedded in the semantic search engine SEMEDICO, for the visualization of complex semantic annotations, namely semantic relations among named entities (such as protein-protein interactions) automatically extracted from biomedical publications.

Keywords: data visualization, graphical user interface, semantic metadata, life sciences

1 Introduction

The biomedical domain has several large-sized, conveniently searchable document stores that are composed of bibliographic data, including abstracts (e.g. MEDLINE)[1] or even full text documents (e.g. PUBMED CENTRAL (PMC)).[2] These collections incorporate up to several millions of publications, e.g. in 2012 more than 21M documents in MEDLINE. Although manually supplied keywords are available both for MEDLINE and PMC documents, the vast majority of the textual bodies come without (further) semantic annotation and, thus, remain un(der)structured in terms of semantic metadata.

Increasingly, these repositories undergo computational text analysis to provide semantic metadata automatically. This includes keyword or keyphrase-style data from automatic indexing systems, e.g. the MEDICAL TEXT INDEXER (MTI) for indexing MEDLINE [1]. As far as information extraction is concerned, named entity taggers recognize instances of semantic classes (e.g. all types of proteins or

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.ncbi.nlm.nih.gov/pmc/

diseases in a running text, which can be expressed, e.g. as *Protein(Interleukin-6)* or *Disease(gastritis)*).

More recently, a new, more complex type of RDF-style metadata has found wider attention, namely semantic relations extracted from unstructured publications where several named entities are linked to form an empirically valid statement (e.g. "protein Interleukin-6 interacts with protein Interleukin-2", or formally, *Interact(Protein(Interleukin-6), Protein(Interleukin-2))*). As these automatically extracted data are getting more and more complex (a semantic relation, e.g. links at least two named entities (such as *Protein(Interleukin-6)* and *Protein(Interleukin-2)* from the example above) via a relator (a predicate name such as *Interact*) and their number is increasing dramatically, standard textual modes for displaying the results of searches on these document sets become inappropriate.

Meanwhile evidence has been gathered that graphical representations are useful for exploring large amounts of relational information [2]. Therefore, we propose an interactive graphical front-end for the visualization of semantic annotations of publications from the life sciences, which is part of the latest prototype of the semantic search engine SEMEDICO.[3] We focus, in particular, on the visualization of relations indicating protein-protein interactions (so-called PPIs), a highly relevant type of relational information for molecular biologists.

2 Search for Biological Information

There are two main types of information available for biological researchers. On the one hand, tons of biological facts are contained in publications, yet due to their verbal format in a (from the perspective of computers, at least) more or less unstructured form. Standard document retrieval, as illustrated by the PUBMED search engine, provides manually assigned index terms (based on the work of human indexers and support tools such as MTI) to add semantic meta data to otherwise unstructured documents. On the other hand, there exist several hundreds of highly specialized biological fact databases (for a recent survey, see [3]). Searchable content is entered into these repositories in a highly structured format (adhering to relational database schemata) by human curators based on their understanding of the contents of articles from selected journals. This procedure yields high-quality structured semantic meta data but is too slow to keep up with the exponential growth of published literature [4], demanding the provision of automatic means for text analysis from a second angle.

Among the many systems trying to improve on PUBMED's search facilities only a few of them are comparable to SEMEDICO. Perhaps closest in terms of scope and functionality is the semantic search engine FACTA+ [5]. FACTA+ sorts search results into several columns representing co-occurring concepts, listing e.g. *cancer* as top co-occurrence in the *Disease* column when searching for *death*. Furthermore, FACTA+ has rudimentary support for searching for PPIs,

[3] `http://www.semedico.org/`

marking whole documents as being relevant for an interaction between two proteins and using a treemap to visualize likely connections.

Visualization is commonly used in the biomedical community to help scientists understand complicated interaction patterns, e.g. via pathway diagrams (see, e.g. [6]). Visualization combined with proper information filtering can greatly improve the presentation of search results [7]. A particularly prominent piece of visualization software in the biomedical domain is CYTOSCAPE.[4] CYTOSCAPE WEB [8] is an online reimplementation of CYTOSCAPE's basic functions, used, e.g. by the highly relevant INTACT PPI database.[5]

To the best of our knowledge no current system, except SEMEDICO, combines automatic relation extraction with a CYTOSCAPE-based visualization tool.

3 Semantic Search Engine Semedico

3.1 Semedico's Architecture

SEMEDICO is a semantic search engine developed at our lab. It utilizes state-of-the-art NLP technology for content-oriented text analytics in a pipeline of analysis engines (including tokenizers, lemmatizers, POS taggers, etc.). This pipeline (see Fig. 1) is based on the *Unstructured Information Management Architecture* (UIMA),[6] which eases the integration of components by providing a central object-based data structure.

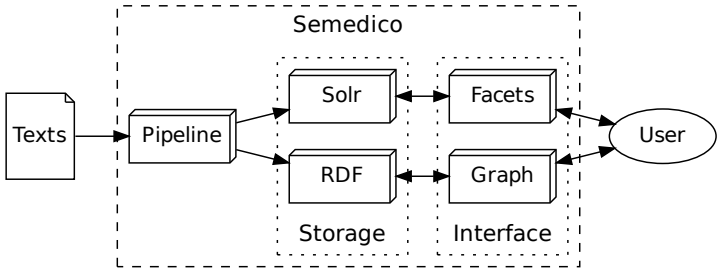

Fig. 1. SEMEDICO's system architecture.

Of particular importance are SEMEDICO's core semantic analysis engines. They encompass a series of named entity taggers, e.g. for genes/proteins (GENO), species, diseases, chemicals, and a suite of relation extractors (JREX), which account for various types of PPIs. GENO performs on a par with the world's best protein tagger based on the BIOCREATIVE II benchmark [9], while JREX [10] outperforms any other PPI extraction tool in the U-COMPARE server [11]. A typical output of JREX consists of an interaction type, e.g. *Upregulates* and two proteins marked for their role in the interaction, e.g. the interaction-causing protein (*agent*) and the resultant protein (*patient*).

[4] `http://www.cytoscape.org/`
[5] `http://www.ebi.ac.uk/intact/`
[6] `http://uima.apache.org/`

We used JReX to extract some ten-thousand semantic relations from approximately 100K Medline abstracts. Originally, Semedico's search technology infrastructure was solely based on Solr,[7] so that documents were indexed by fields containing, e.g. the author's name or a disease found in the text. Since such an index is not intended for storing networks of relations, we combined it with a RDF triplestore. RDF triples, consisting of two entities and a relation, are a natural choice for the storage of PPIs extracted by JReX. Additional metadata, e.g. the unique identifier for the article a PPI was extracted from, are stored with the PPI triple by reifying it. Thus, Solr index-based search and RDF are combined in Semedico—the index provides articles relevant for a specific query, whereas the triplestore serves as back-end for the visualization.

3.2 Interfaces for Semedico

Semedico combines two different modes of interaction to search for and navigate in automatically extracted data. The first mode is a facet-based interface for searching articles as shown in Fig. 2.

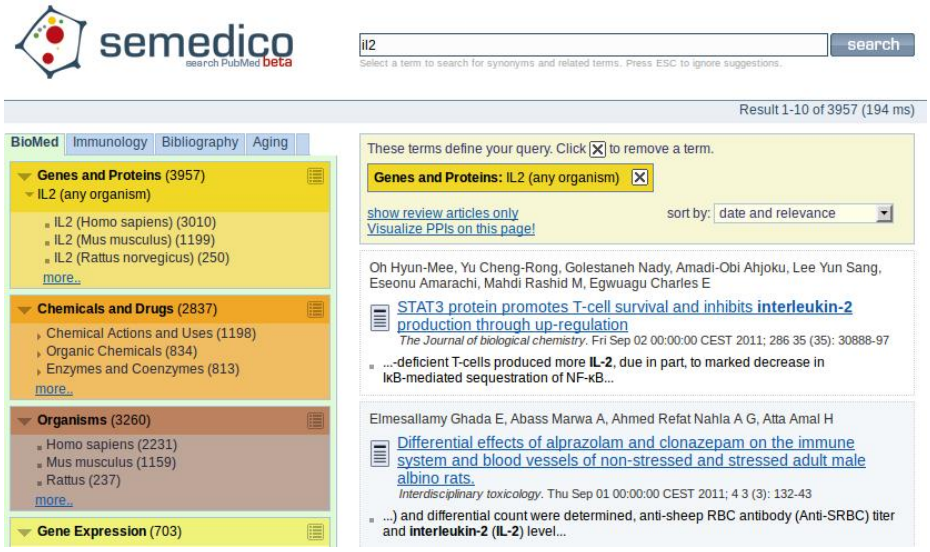

Fig. 2. Semedico' facet-based interface: results for querying for the protein *il2*.

Users can enter queries in a (auto-completing) search box. Queries can use both simple keywords (addressing automatically recognized named entities) or a limited syntax (i.e. *protein Interaction-type protein*) for addressing more complex semantic relations. This interface uses taxonomic information (derived from the MeSH,[8] an authoritative biomedical terminology used for indexing Medline) to filter search results. Queries can be refined by (de-)selecting query terms or

[7] `http://lucene.apache.org/solr/`
[8] `http://www.ncbi.nlm.nih.gov/mesh`

choosing one of several meanings in ambiguous input. This mode also provides a list of results, including highlighted matches for the query and links to the abstract of each article (see lower right half of Fig. 2).

The second mode is graph-based, displaying proteins as nodes and their interactions as edges (see Fig. 3). This mode is available by clicking a link, both while viewing the search results list or while viewing a particular article. Selecting a node displays a link to this protein's UniProt[9] entry (the standard database for protein sequences and functions) at the bottom of the window display. The user can then re-center the graph around a single node to explore all interactions of that protein by clicking that node. Edges are selectable too, triggering the appearance of a link to the abstract that describes this interaction verbatim.

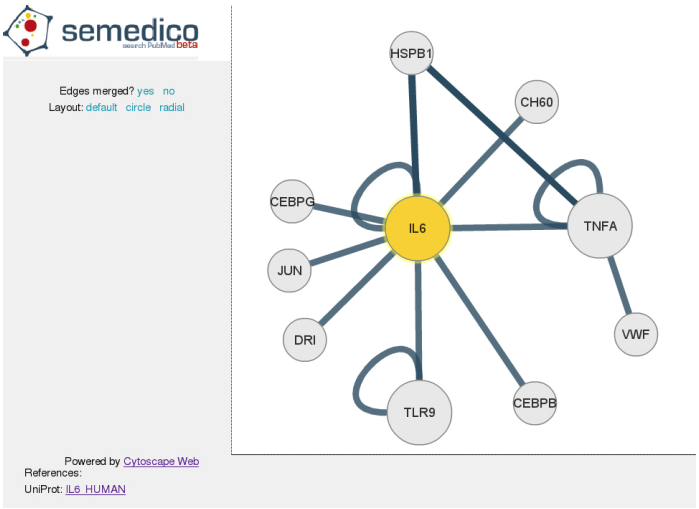

Fig. 3. Graph visualization for PPIs involving *Interleukin-6*.

4 Discussion

We described an online visualization tool for the semantic search engine Semedico that allows users to explore protein interactions found in a document in a comprehensible way. Our prototype improves on the state of the art by combining automatic PPI extraction with graph-style visualization and interaction. The visualization mode offers features not yet common in PPI databases like IntAct, e.g. dynamically reloading interaction graphs without issuing a new query and using edges as links to articles describing that interaction. Open issues of particular importance include the provision of an option to download PPI networks as files for viewing in Cytoscape, directing the generated RDF to other resource collections, e.g. participating in Linked Open Data [12], or combining it with OWL [13] for reasoning tasks. Last but not the least, we have to conduct a usability study, similar in spirit to the one we already carried out for the facetted search mode [14], to find out whether the implemented visualization features are useful for biologists at all. The biggest challenge might be to focus on more

[9] `http://www.uniprot.org`

sophisticated semantic filters, an area where human-curated databases benefit greatly from expert judgments and expertise [7].

Acknowledgments. This work is funded by a grant from the German Ministry of Education and Research (BMBF) for the *Jena Centre of Systems Biology of Ageing* (JenAge) (grant no. 0315581D).

References

1. Aronson, A.R., Mork, J.G., Rogers, W.J., Lang, F.M., Névéol, A.: The NLM Medical Text Indexer: A Tool for Automatic and Assisted Indexing. A Report to the Board of Scientific Counselors, April 10, 2008. Technical report, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services (2008)
2. Gehlenborg, N., O'Donoghue, S.I., Baliga, N.S., Goesmann, A., Hibbs, M.A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., *et al.*: Visualization of *omics* data for systems biology. Nature Methods **7** (2010) S56–S68
3. Galperin, M.Y., Fernández-Suárez, X.M.: The 2012 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Research **40**(Database Issue) (2012) D1–D8
4. Baumgartner, W.A., Jr., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.: Manual curation is not sufficient for annotation of genomic databases. Bioinformatics **23**(13) (2007) i41–48
5. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., Ananiadou, S.: Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics **27**(13) (2011) i111–i119
6. Klukas, C., Schreiber, F.: Dynamic exploration and editing of Kegg pathway diagrams. Bioinformatics **23**(3) (2007) 344–350
7. Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., Cesareni, G.: Mint, the Molecular Interaction Database: 2009 update. Nucleic Acids Research **38**(Database Issue) (2010) D532–D539
8. Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q., Bader, G.D.: Cytoscape Web: an interactive Web-based network browser. Bioinformatics **26**(18) (2010) 2347–2348
9. Wermter, J., Tomanek, K., Hahn, U.: High-performance gene name normalization with GeNo. Bioinformatics **25**(6) (2009) 815–821
10. Buyko, E., Faessler, E., Wermter, J., Hahn, U.: Syntactic simplification and semantic enrichment: trimming dependency graphs for event extraction. Computational Intelligence **27**(4) (2011) 610–644
11. Kano, Y., Björne, J., Ginter, F., Salakoski, T., Buyko, E., Hahn, U., Cohen, B.K., Verspoor, K., Roeder, C., Hunter, L.E., *et al.*: U-Compare bio-event meta-service: compatible BioNLP event extraction services. BMC Bioinformatics **12**(481) (2011)
12. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data: the story so far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22
13. Lacy, L.W.: Owl: Representing Information Using the Web Ontology Language. Trafford Publishing (2005)
14. Schneider, A., Landefeld, R., Wermter, J., Hahn, U.: Do users appreciate novel interface features for literature search?: A user study in the life sciences domain. In: SMC'09 – Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics. (2009) 2062–2067