Bi^{2}SoN – A Digital Library for Supporting Biomedical Research

Benjamin Köhncke
L3S Research Center
Appelstrasse 9a
30167 Hannover
Germany
koehncke@l3s.de

Sascha Tönnies
L3S Research Center
Appelstrasse 9a
30167 Hannover
Germany
toennies@l3s.de

Wolf-Tilo Balke
IFIS TU Braunschweig
Mühlenpfordtstrasse 23
38106 Braunschweig
Germany
balke@ifis.cs.tu-bs.de

ABSTRACT

In the domain of biology a huge amount of different data sources is available. Therefore, information gathering and searching are challenging tasks. To avoid a manual assessment of all relevant data sources, their knowledge has to be integrated. The presented system focuses on all aspects needed for suitable data integration and retrieval for domain experts from the field of biology. The knowledge from different data sources is combined and further used for, e.g. synonym enrichment of the query term. The resulting prototype was presented to a group of domain experts who confirmed that the system delivers suitable results supporting the scientists by their literature search.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval – *Information Search and Retrieval.*

General Terms

Management, Design, Human Factors

Keywords

Architecture, Biology, Personalization, Digital Library

1. INTRODUCTION AND BACKGROUND

Nowadays, the directed information search is usually performed via the Internet. Independent of the field of interest a huge amount of online accessible information is available. Especially in the field of biology searching for relevant data is a challenging task. Biology is a wide domain with many different working areas. Each area has special data sources where the required knowledge is stored. In total there are more than 1000 different data sources storing biological knowledge.

Today the search for information in biology is performed by manually requesting all required data sources and manually combining the retrieved information. The combined information is subsequently used for a literature search in specialized digital libraries. The most widely-used library for the field of biomedical and life sciences information is MEDLINE (Medical Literature Analysis and Retrieval System Online) covering more than 21 million articles from around 4500 different journals. It is compiled by the United States National Library of Medicine (NLM) and freely accessible via the PubMed interface.

The strength and weaknesses of commonly used search portals for biomedical information retrieval are compared in [1], namely PubMed, Web of Science, Scopus, and Google Scholar. The authors compared the content coverage and practical utility of these sources. They used example keyword searches to evaluate the usefulness and specific published articles to evaluate their utility in performing citation analysis. Whereas PubMed and Google Scholar offer free access to their search platform, Scopus and Web of Science are commercial products requiring access fees. The evaluation showed that PubMed is up-to-date due to frequent updates, reliable and easy to use. With GoPubMed [2] it exists a portal allowing users to explore PubMed search results with a hierarchically structured vocabulary for molecular biology, the Gene Ontology [3]. In GoPubMed an overview of the document abstracts according to the Gene Ontology is given. Therefore, the user is able to navigate through the abstracts by category. Furthermore, also more general ontology terms are shown, related to the original query that do not occur directly in the abstracts.

However, all approaches have in common that they only index the terms directly occurring in the documents. But most of the query terms have several synonyms frequently used in different documents. In [4] it was mentioned that for many proteins also different names were used even within the same article. Without taking these synonyms into account a huge amount of relevant literature is not included in the result set. Another aspect is that PubMed does not support a suitable ranking. Usually the result set is ordered by showing the newest documents first. Most of the queries deliver huge results sets making it impossible for the user to find the most important documents without a suitable ranking.

Furthermore, the user's relevance sensation is usually not only query dependent, but also related to his implicit background knowledge and the task he is currently working on. This knowledge cannot be expressed in the query. To summarize, the following additional aspects have to be considered in a retrieval system for biological literature:

- Query expansion using synonyms
- Suitable ranking functions
- Personalized access to information sources

In the course of the Bi^{2}SoN project we will focus on these aspects to build an architecture supporting biomedical researchers with their literature search. Our aim is to support the scientist by

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

offering a portal with direct access to all required data sources. Beside the literature research this also includes an analysis of the required data sources for query expansion. We did a survey with domain experts from the field of infection research and found out that depending on the task the scientist is working on the different knowledgebases for query enrichment differ. We will introduce a framework integrating the knowledge from the required data sources by taking the search task and implicit knowledge of each individual user into account.

2. Bi^{2}SoN ARCHITECTURE

During the project we work together with experts from the Helmholtz Centre for Infection Research (HZI). For the area of infection research it is important to understand bio-chemical reactions in cells. Therefore, a lot of mass-spectrometric experiments are accomplished leading to huge amounts of experimental data. In a first step, this data is analyzed using statistical approaches. An even more important step is the comparison of these results with other approaches known in literature. For example, it is possible that the same reactions are also known from other organisms. Hence, after the relevant entities have been determined from the experimental data, they are further used for literature search. Figure 1 gives an over view of our architecture.

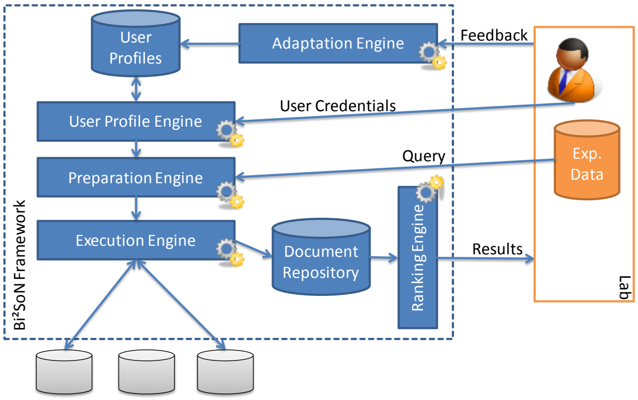
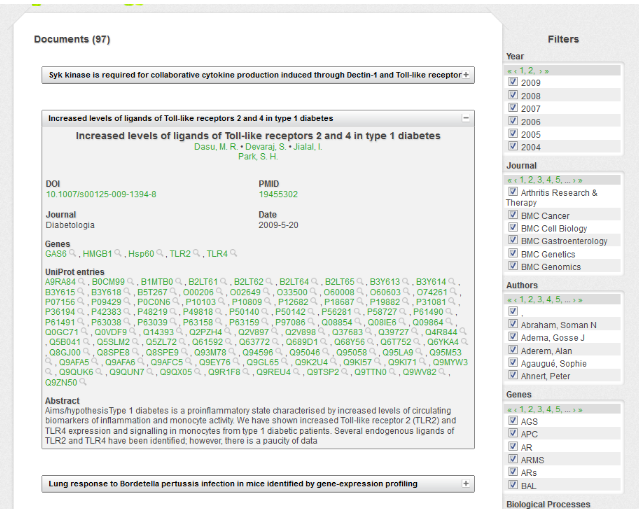


Figure 1. Bi^{2}SoN architecture

The scientist analyzes the experimental data and determines the interesting entities. He identifies himself to the Bi2SoN system and submits a query term out of these entities. For each user a user profile is stored in our repository including information about preferred data sources and working tasks. The user profile engine is responsible for managing the user profiles. The profile and the query is analyzed by the preparation engine which is responsible for choosing adequate data sources. The resulting service set is transmitted to the execution engine which requests the included data sources. Each data source delivers information related to the query entity. The execution engine analyzes and merges the data from each data source and finally transmits the enriched query to the document repository. In a last step a suitable ranking function is chosen taking the user preferences defined in the profile into account. The ranked result set is delivered to the user who has the choice to give a relevance feedback to the system. The user feedback is used to adapt information in the profile, e.g. if the precision of the top-x documents is low another ranking function has to be used for this user.

We implemented the introduced architecture in a first prototype. Our document repository includes around 220000 documents from the PubMed Central corpus. Furthermore, we used external knowledgebases to extract all genes and pathways included in each document. We combined the information from several sources, e.g. UniProt and KEGG, to allow for automatic synonym retrieval. Figure 2 shows a screenshot of our system which we presented to the domain experts from the HZI. The feedback was very positive since the system already retrieves good search results due to the integration of different data sources and synonym enrichment.



Documents (97)

Syk kinase is required for collaborative cytokine production induced through Dectin-1 and Toll-like receptor +

Increased levels of ligands of Toll-like receptors 2 and 4 in type 1 diabetes −

Increased levels of ligands of Toll-like receptors 2 and 4 in type 1 diabetes
Dasu, M. R. • Devaraj, S. • Jialal, I.
Park, S. H.

DOI
10.1007/s00125-009-1394-8

PMID
19455302

Journal
Diabetologia

Date
2009-5-20

Genes
GAS6 , HMGB1 , Hsp60 , TLR2 , TLR4

UniProt entries
A9RA84 , B0CM99 , B1MTB0 , B2LT61 , B2LT62 , B2LT64 , B2LT65 , B3Y613 , B3Y614 , B3Y615 , B3Y618 , B5T267 , O00206 , O02649 , O33500 , O60008 , O60603 , O74261 , P07156 , P09429 , P0C0N6 , P10103 , P10809 , P12682 , P18687 , P19882 , P31081 , P36194 , P42383 , P48219 , P49818 , P50140 , P50142 , P56281 , P58727 , P61490 , P61491 , P63038 , P63039 , P63158 , P63159 , P97086 , Q08854 , Q08IE6 , Q09864 , Q0GC71 , Q0VDF9 , Q14393 , Q2PZH4 , Q2V897 , Q2V898 , Q37683 , Q39727 , Q4R844 , Q5B041 , Q5SLM2 , Q5ZL72 , Q61592 , Q63772 , Q689D1 , Q68Y56 , Q6T752 , Q6YKA4 , Q8GJ00 , Q8SPE8 , Q8SPE9 , Q93M78 , Q94596 , Q95046 , Q95058 , Q95LA9 , Q95M53 , Q9AFA5 , Q9AFA6 , Q9AFC5 , Q9EY76 , Q9GL65 , Q9K2U4 , Q9KI57 , Q9KI71 , Q9MYW3 , Q9QUK6 , Q9QUN7 , Q9QX05 , Q9R1F8 , Q9REU4 , Q9TSP2 , Q9TTN0 , Q9WV82 , Q9ZN50

Abstract
Aims/hypothesisType 1 diabetes is a proinflammatory state characterised by increased levels of circulating biomarkers of inflammation and monocyte activity. We have shown increased Toll-like receptor 2 (TLR2) and TLR4 expression and signalling in monocytes from type 1 diabetic patients. Several endogenous ligands of TLR2 and TLR4 have been identified; however, there is a paucity of data

Lung response to Bordetella pertussis infection in mice identified by gene-expression profiling +

Filters

Year
« ‹ 1, 2, › »
☑ 2009
☑ 2008
☑ 2007
☑ 2006
☑ 2005
☑ 2004

Journal
« ‹ 1, 2, 3, 4, 5, ... › »
☑ Arthritis Research & Therapy
☑ BMC Cancer
☑ BMC Cell Biology
☑ BMC Gastroenterology
☑ BMC Genetics
☑ BMC Genomics

Authors
« ‹ 1, 2, 3, 4, 5, ... › »
☑ ,
☑ Abraham, Soman N
☑ Adema, Gosse J
☑ Aderem, Alan
☑ Agaugué, Sophie
☑ Ahnert, Peter

Genes
« ‹ 1, 2, 3, 4, 5, ... › »
☑ AGS
☑ APC
☑ AR
☑ ARMS
☑ ARs
☑ BAL

Biological Processes

Figure 2. Bi^{2}SoN portal screenshot

The user has the ability to filter the result set using bibliographic and semantic facets. The description of each document includes, among others, all genes and UniProt entries.

3. CONCLUSION

In the field of biology a huge amount of different data sources is available. To allow for high quality document retrieval an integration of this knowledge is essential. We have shown in a first prototype that the integration of different sources for synonym enrichment already leads to pleasant retrieval results. For future work we plan to further investigate the definition of the user profile. Beside the already included information about preferred data sources and working tasks, it would be interesting to model the implicit knowledge of a biologist. Usually he has an exact perception of what he is interested in which cannot be expressed in a query. Moreover, it is important to invent personalized ranking functions to further improve the user's search experience.

4. REFERENCES

[1] M. E. Falagas et al., "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses.," *The FASEB journal*, vol. 22, no 2, 2008.

[2] A. Doms, M. Schroeder, "GoPubMed: exploring PubMed with the Gene Ontology.," *Nucleic acids research*, vol. 33, no Web service issue, 2005.

[3] M. a Harris et al., "The Gene Ontology (GO) database and informatics resource.," *Nucleic acids research*, vol. 32, no Database issue, 2004.

[4] L. Hirschman et al., "Accomplishments and challenges in literature data mining for biology.," *Bioinformatics (Oxford, England)*, vol. 18, no. 12, 2002.