The LAILAPS Information Retrieval Portal as Scalable and Integrative Database Query Endpoint

Matthias Lange[1], Jinbo Chen[1], Daniela Borck[2], Christian Colmsee[1], Klaus Hippe[2], Benjamin Kormeier[2], Stephan Weise[1], Uwe Scholz[1]

[1]Research Group BIT, IPK Gatersleben, Corrensstr. 3, D-06466 Gatersleben Germany
[2]Bielefeld University, Bioinformatics Department, PO Box 10 01 31, D-33501 Bielefeld, Germany

Summary

Retrieval and citation of primary data is the important factor in the approaching age of "data science". Solving the challenge of building a flexible but homogeneous bioinformatics information retrieval infrastructure to access and query the world life science databases is a crucial factor for an efficient building bioinformatics infrastructure.

LAILAPS search portal was designed to provide the look and feel of a web search engine. To support a platform independent implementation and scalable service, we decided to use a JAVA 3-tier web application. The frontend is a web application that supports a keyword based search, a browser for relevance ordered query result, and a data browser to inspect and rate matching data records. We use a relevance probabilistic prediction model using neural networks. To consider the fact that data relevance is highly subjective to the user of an information retrieval system, we support user or use case specific neural networks. In order to train a particular neural network, we presented the network a reference set of 9-dimensional feature vectors as input and for the output layer the related manual curated relevance score.

LAILAPS provides a central entry point to institutes or projects databases. The advantages of LALAPS compared to other database search portals in life science is the dynamic adjustable relevance ranking in combination with customizable system configurations, like synonym lists and relevance scoring keywords, and an embedded feedback and training system. The system is available as a wizard style installation package and enables the set-up of user customized search portals and is available under the URL http://lailaps.ipk-gatersleben.de

1 Introduction

Nowadays, life science institutions provide web applications or web services to query the content of their hosted databases. Those data access services integrate more or less all available databases in one or more services. The services range from web service APIs, flat file data downloads to web frontends. Each published life science database institution defines its own data publication strategy or, potentially problematic, database specific heterogeneous solutions. We suggest overcoming such isolated data query end-points without interfering the individual database infrastructure using information retrieval (IR). In the context of the described scenario, IR means the retrieval of relevant data from distributed and heterogeneous scientific data sources.

Because IR is an active bioinformatics research area, there are several strategies. Formula based database frontends or complex information systems are the intuitive and commonly used way to bring data to the scientist's desktop. But picking the most relevant database

entries out of thousands of database records is a time consuming and frequently un-deterministic task. In this context, keyword search refined by criteria like who published, in which journal, for which organism, evidence scores, surrounding keywords etc. matter. This approach is general known as Boolean Search.

Methods supporting an automatic relevance ranking where developed in the last decades. Established models for relevance ranking are Boolean Searches, the Vector Space Model (VSM) or the Term Frequency - Inverse Document Frequency (TF-IDF) model. It is frequently used in information retrieval. As ranking model for interlinked data, the pagerank algorithm was developed. Its hypothesis is: A high hyperlink in-degree of a page means high popularity and high popularity means high relevance. Another approach is the probabilistic relevancy ranking, whereby weight parameters for the relevance of database fields and word combinations have to be predefined. In combination with a user feedback system, the probabilistic approach shows promising ranking performance.

In this contribution we apply the LAILAPS search engine [1] as a system that combines a range of well discriminating database relevance features within a probabilistic model under consideration of user specific relevance profiles. In this abstract we present the application of the LAILAPS search engine as customizable information retrieval portal for different institutions with different kind of use cases. Based on a use case, we demonstrate common application scenarios and the influence of user specific relevance profiles to the relevance prediction performance. The LAILAPS software is available under the URL http://lailaps.ipk-gatersleben.de

2 Result and Discussion

The intention of LAILAPS information retrieval portal is the provision of an information retrieval infrastructure that meets the requirements of the “data science” age and offers a re-usable information retrieval platform for data exploration, data citation, and data publication. Like a shopping cart in a shopping mall, the aim of the project is to implement a search engine and a data cart, which bridge the information retrieval gap over hundreds of non-integrated databases. We suggest a loosely over metadata linked data network in combination with a simple, HTTP-based data access. This is the most non-invasive and therefore most cost saving way to build for existing databases and bioinformatics information systems an IR infrastructure. The LAILAPS approach is to provide ad-hoc installable individual IR portals with a minimum effort of configuration and no effort for database integration, interface implementation or frontend coding.

2.1 The LAILAPS Search Engine

LAILAPS is an information retrieval system for life science databases. The system was designed to provide the look and feel of a web search engine. To support a platform independent implementation and scalable service, we decided to use a JAVA 3-tier web application.

The frontend is a web application that supports a keyword based search, a browser for relevance ordered query result, and a data browser to inspect and rate matching data records. The included feedback system enables the user to train the relevance prediction system with individual relevance ratings. Management and administration pages complete the frontend.
The search engine logic is implemented as JAVA RPC service. It offers the following functions: *query parsing, synonym expansion, query suggestion, text search of query keywords, feature extraction, relevance ranking of matching database records, result*

delivery, relevance feedback collection, training of neural networks for relevance prediction, user, and session management, system configuration.

The backend database stores the imported life science databases in the flexible NO-SQL approach as key-value pairs, whereas the key means the database attribute. This enables the seamless inclusion of any kind of data source that supports a flat CSV export, which is expected to be supported or ad-hoc implementable by the majority of databases.

2.2 Portals using LAILAPS Search Engine

We will demonstrate the benefit of Search Portals for keeping data accessible and retrievable by providing one unique data query retrieval hub. The aim of the LAILAPS research team is to provide a customizable, ad-hoc installable software system to building customized database search portals. The offered software supports data centers and primary data archives that like to provide a single point of service to their databases and information systems. In order to provide a proof of concept, three LAILAPS portals where installed (see table 1).

Search Portal	URL	#databases	#indexed records	Data domain
UniProt	http://pgrc.ipk-gatersleben.de/lailaps	2	1.215.553	protein annotation
IPK Databases	http://pgrc.ipk-gatersleben.de/lailapsipk	5	287.252	EST annotation, expression, plant phenotype, metabolite, metabolism
DAWIS-M.D. [2,3]	http://pgrc.ipk-gatersleben.de/lailaps_dawis	11	9.461.855	disease, ontology, drug, protein metabolism, gene functions, compounds

Table 1: overview to the current installed LAILAPS Search Portals. More detailed information is available using the link "Which databases are indexed?" available at the portals home page.

The first for UNIPROT databases as example for public data center. The databases where downloaded as flat files from the UniProt FTP-server and transformed to LAILAPS CSV import format. The second portal was installed for IPK in-house databases. Here, the situation was a high number of distributed and isolated information systems based on ORACLE DBMS. Here, we generated the LAILAPS import files using proprietary SQL statements. At last we installed a portal for the DAWIS-M.D. data warehouse of integrated life science databases, developed at Bielefeld University, Germany [2,3]. They provided a CSV version of the warehouse content.

2.3 Benchmarking the Relevance Ranking

The key feature of LAILAPS is the relevance prediction. We use a relevance probabilistic prediction model using neural networks. To consider the fact that data relevance is highly subjective to the user of an information retrieval system, we support user or use case specific neural networks. In order to train a particular neural network, we presented the network a reference set of 9-dimensional feature vectors as input and for the output layer the related manual curated relevance score. Each set consists of use case specific queries and the relevance rating of the query results. In [1] we presented a training set for plant metabolic data. Using this set, we trained the default neural network for the guest user. To present the effect of use case specific ranking, we subsequently present the results of training the DAWIS-M.D. portal based on signal transduction data.

The signal transduction is a very important and complex process in an organism, which regulates many biological processes like cell proliferation or gene transcription. A lot of enzymes and other biological elements are involved. Some of these substances take a major role in the signal transduction. Key substances in the signal transduction are Heparan Sulfate Proteoglycans (HSPGs). HSPGs are involved in several protein-protein interactions like the regulation of the binding from various growth factors. This controls the cell growth and cell differentiation. Exact information retrieval about HSPGs and their correlation is difficult, because many synonyms and homonyms in the databases complicate the exact searching. Also a lot of databases provide incomplete data sets, so that important information is not available. Another critical point is the identification of interactions between other biological elements. In discussion with our biochemical experts, we have made several test queries about HSPGs, HS (Heparan sulfate) and SULF1/2 and rate each relevant entry. Table 2 shows 9 search phrases and their prediction performance before and after the training.

		default ranking network			signal transduction network		
query	**result size**	**RMSE**	**exact match rate**	**max error**	**RMSE**	**exact match rate**	**max error**
"heparan sulfate proteoglycan" sulfatase	9	0.49	0.11	0.7	0.24	0.67	0.70
"heparan sulfate proteoglycan" sulf1	6	0.57	0.33	0.7	0.35	0.67	0.60
"sulfatase 1"	7	0.54	0.00	0.6	0.24	0.29	0.60
"sulfatase 2"	7	0.56	0.00	0.6	0.00	1.00	0.00
sulf-1	5	0.49	0.50	0.6	0.27	0.80	0.60
sulf1	7	0.53	0.29	0.7	0.32	0.57	0.60
sulf1 sulf2	7	0.46	0.14	0.6	0.08	0.29	0.10
sulf-2	5	0.38	0.20	0.6	0.41	0.60	0.70
sulf2	10	0.61	0.00	0.7	0.27	0.70	0.60

Table 2: The nine queries where selected to cover a wide spectrum of modified sulfates of HSPGs. We searched for gene and protein names in several spelling variants and keyword combinations. The column "RMSE" is the rooted mean square error as estimator of true and predicted relevance. The "max error" is the maximum difference of true and predicted relevance. The "exact match rate" is the rate of the entries with exact match between predicted relevance and the manual rated.

In average we observed after the use case specific training an error rate of 0.24 after compared to 0.51 before use case specific training. This is a decrease by factor 2.1. The average rate of exact matches increase by factor 3.6 from 17% to 62%. The maximal absolute error is in both cases not satisfying. This is caused by outliers and the small number of training data. Using the feedback system of the LAILAPS frontend the training was done in one day work. Furthermore, an additional adjustment of attribute and database scores we expect a better relevance prediction. For the planed comprehensive study of HSPGs we will also increase the training set and do a peer curation of entry relevance. Furthermore, we work on extension the synonyms and keyword list for LAILAPS DAWIS-M.D. portal.

To summarize, by providing portal with the support of individual relevance prediction profiles, the information retrieval has a higher quality and bridge isolated and heterogeneous databases in a non-invasive way. Additionally, the manual information retrieval by query

refinement is no longer time-consuming, the scientist can now efficient, with a reproducible quality, pick the data "nuggets" in the data universe. We finally conclude that information retrieval may help to increase the visibility and reusability of scientific data.

References

[1] M. Lange, K. Spies, C. Colmsee, S. Flemming, M. Klapperstück and U. Scholz. The LAILAPS Search Engine: A Feature Model for Relevance Ranking in Life Science Databases. Journal of Integrative Bioinformatics, 7(3):e118, 2010.

[2] K. Hippe, B. Kormeier, T. Töpel, S. Janowski and R. Hofestädt: DAWIS-M.D. - A Data Warehouse System for Metabolic Data. GI Jahrestagung (2) 2010: 720-725

[3] T. Töpel, B. Kormeier, A. Klassen, and R. Hofestädt. BioDWH: A data warehouse kit for life science data integration. Journal of Integrative Bioinformatics, 5(2):93, 2008.