Community curation for GeneView

Studienarbeit

Exposé

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

bearbeitet von: Alexander Konrad

Betreuer: Prof. Dr. Ulf Leser, Philippe Thomas

Datum: 19.02.2013

1 Motivation

The latest discoveries of diseases and their diagnosis or treatments have been mostly published in scientific literature. The fast growth of published biomedical articles led to a strong ambiguity of disease names meaning a traditional keyword-based search for biomedical articles will not lead to satisfying results [DL12]. This problem does not only exist for the terms of diseases, it's a problem for most names of biomedical objects like genes or chemicals too [TSV+12]. As long as data grows exponentially, novel Biomedical Informatics approaches and tools are needed to retrieve the data [FMM07]. Efficient search tools are crucial for biomedical researchers to keep abreast of the biomedical literature relating to their own research [DM+09].

GeneView is such a tool. It was created as a comprehensively annotated version of PubMed articles and abstracts. PubMed[1] is perhaps the most popular information retrieval tool in biomedicine based on MEDLINE[2] repository [S10]. GeneView is a semantic search engine for PubMed using a multitude of state-of-the-art text mining tools for recognizing instances of ten entity classes and protein-protein interactions (PPI) [TSV+12]. GeneView provides possibilities to search for scientific biomedical articles with a number of features.

Systems like GeneView use automatic methods to annotate biomedical articles automatically. But the precision and recall of the used algorithms still left plenty of non-detected (false-negatives) or wrongly detected (false-positives) entities. As example the gene name recognition uses the tool GNAT which has a precision of 82% and a recall of 82% for abstracts and 54/47% respectively for full articles. To close the gaps of not detected entities a manually curation of the results is necessary. Leitner et. al show in [LCA+10] that annotations made by systems and manually achieved annotations through authors or curators could assist each other to improve the overall performance. Currently there is no possibility to curate annotations manually in GeneView.

GeneView visualizes articles/abstracts with entity highlighting. Highlighting annotated entities is a clever way showing up interesting links between entities. This could be extended by having all information regarding one entity, across all relevant articles, in one place. Many knowledge resources are compiled by manually curated knowledge extracted from biomedical literature and other sources [BB08]. As GeneView has a large repository of information about entities this knowledge can be examined through creating

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

[2] MEDLINE is the U.S. National Library of Medicine's® (NLM) premier bibliographic database that contains over 19 million references to journal articles in life sciences with a concentration on biomedicine. http://www.nlm.nih.gov/pubs/factsheets/medline.html

a new view on the data. Both the need for curation and the desirable new view on the existing data go hand in hand. Having all information about one entity at one place could help the process of curation and having a curated set of data leads to a huge benefit of the quality of the annotations. This work is dedicated to model and implement this new view in GeneView next to developing a model for community curation for annotated entities.

2 Background

This work is based and will be implemented on GeneView. GeneView is a semantic search engine for the Life Sciences. GeneView identifies instances of ten different entity types and three relationship types and indexes approximately 21.4 million abstracts and almost 360.000 full text. To identify these entities and relationships GenView uses a group of state-of-the-art tools. The idea of GeneView is to bundle the best available algorithms into a complex pipeline to analyse every article with different methods to benefit from all individual fortes of the algorithms. GeneView is updated regularly pulling new articles by PubMed. All entities and relationships extracted by the pipeline are stored in a relational database while the articles (or abstracts) are separately stored in Lucene[3] which serves as storage, query and ranking engine. Aggregated information for each entity type is also stored in Lucene after the pipeline has finished [TSL12].

Users can access GeneView through a web interface and query the Lucene index to search for articles. Once a user has sent its query, GeneView provides a list of matching articles sorted by the ranking engines of Lucene (by default this is the publishing date). Each article is shown as a short excerpt with additional information title, authors, released magazine, publishing dates and other facts specific to the article. A click on an article leads to the full text version respectively the abstract. A search for a specific entity will end up in this list as well as a free text search. The result of an search query is document-centric and provides a list of relevant articles from where the user has to pick the articles which look interesting. GeneView provides no possibilities to gain all information about an entity at one place. A way of displaying information cummulated is currenty not implemented. All entities stored in the relational database corresponding to the article are highlighted in a specific color. The highlighting is achieved through different layers, each containing highlighting for precisely one entity. Each layer is an HTML representation of the article. For each entity additional information is provided through a small opening pop-up window when the entity is clicked

[3] http://lucene.apache.org/core/

on. The pop-up shows links to external reference databases and provides more services in the case of proteins or genes. For these entities the pop-up window shows information on pathways and protein-protein interactions the gene/protein participates in. This specific information is aggregated in the background and requested through an AJAX call. Currently there are no curating possibilities provided for GeneView [TSL12].

3 Aims of this work

This work has two main goals. The first one is to achieve a new search method. It should be possible to get all occurrences of an entity over the whole document corpora at once. A list of all sentences containing the searched entity should be provided. The second one is to provide possibilities for community curation. Every sentence should be able to be curated.

4 Approach

4.1 Show all information regarding one entity at once

To find articles the user has to query GeneView through the web interface. Once an article is selected the user sees the annotated version with all annotations highlighted in type-specific colors. All recognized entities provide additional information in an upcoming pop-up window when clicked on. It also provides for every entity the option to search GeneView for articles containing the selected entity. This last point is where this work will start from. It would be desirable to have a special search type available which displays all occurrences of the searched entity in every article in the GeneViews database. The idea is to have all of the already collected knowledge of the entity presented as an extract in the form of specific sentences which contain the entity.

4.2 The ranking options

This new view on an entity will provide for every occurrence of the entity the corresponding sentence. This could require gathering and displaying a rather long list of sentences all containing the searched entity. Thus it is necessary to provide ranking options. The ranking principle should be similar to the ranking of the article list resulting from an article search but should be extended and adjusted. Ranking the sentences is possible on document level as well as on sentence level.

On article level the ranking of the sentences will be sorted by the publishing date of the article. On sentence level the sentences should be sortable by attributes such as the latest curations or the most curated. Additional a free text search should be integrated on sentence level. Only sentences with the respective entity and keyword entered by the user should be displayed. Ranking on NER confidence value: Some named entity recognition tools provide a confidence index which expresses the correctness probability of the found entity. This value could be used as ranking factor. The sentences with the most probable entities would appear on the top, the less probable entities would be display on the bottom of the result list. As the validity of these confidence values are limited and furthermore not available for all entities, this ranking feature should be optional.

4.3 Curate information

Because text mining methods do perform a large number of errors (see above), it is desirable to have an option to curate each sentence individually. The curation possibility will be provided as dop-down box where fixed statements could be selected. Statements should be shown separately while validations should be shown cummulatively. An open question is how to handle different validations of an annotation. What happens if two different curators curate the same entity annotation differently? Or is it only allowed to curate an entity/a sentence uniquley? The information entered by the curator is saved in the database related once to each validated entity in a sentence. GeneView allows a registration for users which is not necessary for just searching articles. Registration should be required to be able to curate. All curations should be public but the curator's name should never be shown. Another question is how to handle a wrong or misleading curation of sentences/entities. How to detect them? How to validate curations? Should there be an option to alert curations to the administration of GeneView? Or should this problem be solved by the community?

A more general point to be discussed is the update behavior of GeneView. There is currently no incremental update mechanism. Instead the whole document corpora gets deleted and rebuild if new documents should be integrated. What should happen to the curated annotations? Would they stay untouched or would they be affected by this update concept?

4.4 Implementation

To implement the new search method it is necessary to analyse the existing code of the web interface. The implementation must fit into the current

environment and must use existing code where it is appropriate. It is necessary to create a search class which is reusable for other entities. First the "gene" entity will be implemented. Afterwards the "chemical" entity will be implemented. The way of implementation should be designed to be easily adopted by other entities. It is desirable to implement the new search method for all entities. The pop-up with augmented information for each entity will be used to link to the new search. The AJAX request for pulling these information from the backend has to be adjusted. For displaying the summarized entity information the concept and methods of highlighting are best applied. This leads to develop a new or advanced virtualization engine.

5 Related work

A large number of scientific publications are concerned with the matter of data curation. The purpose of these papers differs from describing solutions for specific topics to general articles which range from fundamental question regarding curation. Especially in biomedical science curating data is an huge challenge. Rico et al. describe in [PKI+08] WikiPathways. WikiPathways is a curating resource based on MediaWiki open source software and is concerned with biological pathways. They provide an individually adjusted platform to collect and maintain pathway information from the biological community. WikiPathways is open for public participation addressing students as well as senior experts. It contains a custom pathway editing tool which is designed to reduce the barrier that prevents participation in pathway curation. Cariaso and Lennon describe in [CL11] SNPedia. SNPedia is also a wiki-based open access tool designed to curate information about medical, phenotypic, forensic and genealogical associations of DNA variations. The entries are formatted in a systematic way to be read by human or machines. The data is collected both automatically and manually. Entries by users are augmented through periodic updates text mined from public data sources. They describe four levels of data curation where the manually entered data is reviewed: By other users, through semantic warning flags set by Mediawiki templates, through crawling data from outside to augmenting where irregularities are often detected and by reporting the SNPedia content to Promethease[4] personal genome report where it is read by diverse audiences.

On Wiki based systems the question of accounting for the trustworthiness of user generated content rises.The idea of SNPedia to have warning flags automatically set through templates or having a promising concept of

[4]Promethease is a tool to build a report based on SNPedia and a file of genotypes. Customers of testing services [..] can use it to learn more about their DNA" http://snpedia.com/index.php/Promethease

user involvement in detecting irregularities is crucial for curating data. Clearly curation of data in GeneView needs a model for validating the curation. Whereas Howe et al. allege in [HCG+08] that biocuration increasingly lags behind data generation in funding, development and recognition they claim that different parties like authors, journals and curators should begin working together to develop agreements, simplifying the exchange between literature and databases through approved gene symbols or protein accession numbers. They also claim that community-based effort has to be facilitated by researchers and curators and they demand an increasing visibility and support for scientific curation as a professional career. Goble et al. deal in [GSH+08] with curation of web services for data integration in bioinformatics. They state that web services tend to be poorly described, followed by insufficient documentation and describe a set of questions to be solved when discussing curation. They summarize their concerns in an obvious way: "Knowing for whom curation is intended is a guide for what curation is to provide; where it should be done and by whom." In [YBS+12] Yusuf et al. state an interesting point in the behavior of curators in GeneTests. In GeneTests expert authors are recruited to take intellectual ownership of an article of a gene, which leads to an interesting point of the identity of the curators: The are strongly acknowledged and can receive recognition for their intellectual contribution. This could be a hint for only allowing registered users curating data in GeneView. Or maybe there could be a second level curators like the more expert or the more experienced curator could take ownership of specific entities.

6 References

[BB08] Burgun A, Bodenreider O: Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform 2008, 91-101.

[CL11] Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res. 2012;40:D1308–D1312.

[FMM07] Frey LJ, Maojo V, Mitchell JA. Bioinformatics linkage of heterogeneous clinical and genomic information in support of personalized medicine. Methods Inf Med 2007;46 Suppl 1:98-105.

[DM + 09] Dogan I, R, Murray, GC, Neveol, A., Lu, Z. Understanding PubMed user search behavior through log analysis. 2009. Database (Oxford): bap018.

[DL12] Dogan RI, Zhiyong Lu. An improved corpus of disease mentions in PubMed citations. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 91-99.

[GSH + 08] Goble C, Stevens R, Hull D, Wolstencroft K, Lopez R. Data curation + process curation = data integration + science. Brief Bioinform. 2008;9:506–517.

[HCG + 08] Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, et al. Big data: the future of biocuration. Nature. 2008;455:47–50.

[LCA + 10] Leitner F, Chatr-Aryamontri A, Ceol A, Krallinger M, Licata L, Mardis S, Hirschman L, Cesareni G, Valencia A. Enriching Publications with Structured Digital Abstracts: The Human-Machine Experiment. Accepted for publication in Nature Biotechnology, 2010.

[PKI+08] Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6:e184.

[S10] Sarkar IN. Biomedical informatics and translational medicine. J Transl Med. 2010;8:22.

[TSL12] Thomas P, Starlinger J, Leser U. Experiences from Developing the Domain-Specific Entity Search Engine GeneView. 2012.

[$TSV + 12$] Thomas P, Starlinger J, Vowinkel A, Arzt S, Leser U. GeneView: a comprehensive semantic search engine for PubMed. 2012, Nucleic Acids Res. Vol.40(Web Server issue):W585-91

[$YBS + 12$] Yusuf D, Butland SL, Swanson MI, Bolotin E, Ticoll A, Cheung WA, Zhang XY, Dickman CT, Fulton DL, Lim JS, et al. The transcription factor encyclopedia. Genome Biol 13(3): R24.2012