BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1734–1735
doi:10.1093/bioinformatics/btr181

Databases and ontologies

Advance Access publication April 14, 2011

'Sciencenet'—towards a global search and share engine for all scientific knowledge

Dominic S. Lütjohann[1,†], Asmi H. Shah[2,†], Michael P. Christen[2], Florian Richter[2], Karsten Knese[2] and Urban Liebel[2,*]

[1]Institute of Organic Chemistry (IOC) and [2]Institute of Toxicology and Genetics (ITG), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Modern biological experiments create vast amounts of data which are geographically distributed. These datasets consist of petabytes of raw data and billions of documents. Yet to the best of our knowledge, a search engine technology that searches and cross-links all different data types in life sciences does not exist.

We have developed a prototype distributed scientific search engine technology, 'Sciencenet', which facilitates rapid searching over this large data space. By 'bringing the search engine to the data', we do not require server farms. This platform also allows users to contribute to the search index and publish their large-scale data to support e-Science. Furthermore, a community-driven method guarantees that only scientific content is crawled and presented. Our peer-to-peer approach is sufficiently scalable for the science web without performance or capacity tradeoff.

Availability and Implementation: The free to use search portal web page and the downloadable client are accessible at: http://sciencenet.kit.edu. The web portal for index administration is implemented in ASP.NET, the 'AskMe' experiment publisher is written in Python 2.7, and the backend 'YaCy' search engine is based on Java 1.6.

Contact: urban.liebel@kit.edu

Supplementary Material: Detailed instructions and descriptions can be found on the project homepage: http://sciencenet.kit.edu.

Received on February 1, 2011; revised on March 9, 2011; accepted on April 3, 2011

1 MOTIVATION

Most commonly known search engine technologies (Bing, Google) are based on popularity ranking algorithms. However, scientific research has special requirements for search engines that cannot be addressed by popularity ranking in all cases. Special search engines (for example, Scirus (McKiernan, 2005), PubMed, Google Scholar, Web of Science, Scopus) concentrate more on providing content from scientific journals and literature (Falagas *et al.*, 2008).

Other meta search engines cross-link several centralized databases via a single search interface [for example Bioinformatic Harvester (Liebel *et al.*, 2004), EB-eye (Valentin *et al.*, 2010), Entrez (Schuler *et al.*, 1996), Ensembl (Flicek *et al.*, 2010), STRING (Szklarczyk *et al.*, 2010)]. Today's scientific search queries require searching across different data sources that are geographically distributed. Often different data types, like high content screening (HCS) image data or sequence based data (Birney *et al.*, 2007), require special databases that present a challenge to the global search methods mentioned above.

The latest developments in high content/high-throughput screening microscopy (Pepperkok and Ellenberg, 2006) and next-generation sequencing technologies (Metzker, 2010) routinely produce experimental datasets in the terabyte (TB) range resulting in millions of data files. To the best of our knowledge, there is no central database to encompass all experiment datasets due to the fact that large-scale data handling is a challenge for any known data publication platform. Uploading all this data to a centralized database is currently too time consuming and expensive (Schadt *et al.*, 2010). Also, maintaining a centralized infrastructure over the years is costly (Ball *et al.*, 2004). Consequently, it is likely that no single library alone will be able to index the entire science web (Lewandowski and Mayr, 2006). Research strongly benefits from accessible data that provides a valuable resource for comparative and novel studies (Campbell, 2009). Thus, a decentralized search and publishing network that can handle multiple data types at different locations will significantly improve the scientific research process.

2 DESIGN AND IMPLEMENTATION

We designed Sciencenet, a distributed peer-to-peer search engine network that can incorporate many different scientific data types like text, large-scale image datasets (Swedlow and Eliceiri, 2009), DNA/protein sequences (Ansorge, 2009) and mass spectrometry (MS) data (Gstaiger and Aebersold, 2009), which are published on web servers. It facilitates linking search results to other related heterogeneous data sources.

To ensure the scalability of the data space, documents are located via a Distributed Hash Table (DHT) (Balakrishnan *et al.*, 2003). This avoids asking every peer to receive a complete search result. Our DHT rule allows storing index elements for a single search request on several peers. Due to concurrent queries, the more peers contribute, the better the response time gets.

The distributed Sciencenet software platform has the following key elements:

(1) A large-scale index technology capable of handling billions of documents belonging to the scientific web. Based on KIT's 350 000 web pages and currently 6471 known scientific sites in the whitelist,

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

we estimate a total number of over 2 billion documents to be integrated.

Our startup environment for Sciencenet consists of 30 commodity PCs, equipped with 2–24 cores, 4-64 GB of RAM and 500 GB hard disks, each capable of handling 15 million documents, which would just require a total number of about 200 peers for the estimated data space. The operating system is standard Ubuntu 10.04 with Java. This architecture was chosen to mimic a global distributed search engine.

These Sciencenet PCs (peers) are configured to crawl (load and analyze) distinct scientific web sites and import repositories that provide an Open Archive Interface (OAI) (Lagoze *et al.*, 2002). OAI is a standard to import data sources in a fast and structured manner. Currently, 240 million web pages and documents are in the index of our machines. Furthermore, 1 TB of image-based data are available.

The scientific community can easily provide server capacity to expand the index and improve search performance.

(2) A community-driven method to manage the integration of institutional web sites, databases and journals to improve the quality of the scientific search index. Any scientific web site can be submitted by anyone, and registered users can be part of the process to accept these suggestions to support the growth of the index.

(3) A simple 'one-stop' search interface for all users. The Sciencenet web site (http://sciencenet.kit.edu) provides a search portal without installation. The search results are presented along with a domain navigator and a tag cloud to refine the search.

Alternatively, users can download the free open-source Sciencenet-YaCy client software package, allowing them to access the search network from their machines, perform search queries and access published scientific experiment data from others. The result list can be exported via an Application Programming Interface for further processing in external tools.

Due to the preselected index, we consider every search result to be relevant, so pre-computed ranking, like PageRank (Brin and Page, 1998), is not used. The results are ranked using a default 'ranking matrix' consisting of a set of 28 statistical ranking criteria, such as 'word distance' or 'appearance in title' (see Supplementary Material). For each search query, users can customize the values of the ranking matrix with no increase in the overall complexity.

(4) An easy to use software tool that allows data publishing and sharing. Users are able to publish and share their own scientific data or web sites. We provide an example module (the 'AskMe' tool) for non-text based data integration in the downloadable client. The tool handles large-scale image datasets from HCS experiments by providing a dataset preview. All collected meta information is presented in corresponding experiment descriptor files in both human and computer readable form. Hence, we use the embedded Resource Description Framework RDFa (Birbeck and Adida, 2008). This data publication method follows the principle of a Linked Open Data architecture (Berners-Lee, 2006) and is already the foundation for a semantically enriched web (Jensen and Bork, 2010).

3 CONCLUSION AND OUTLOOK

The combination of the technologies mentioned above makes it possible to search thousands of heterogeneous data sources with billions of documents and datasets. Our decentralized peer-to-peer approach overcomes the performance and capacity limitations of centralized data repositories. Ideally, future modules would allow users to rank and comment on search results.

ACKNOWLEDGEMENTS

The authors thank YaCy's open-source community, Björn Kindler for technical support, Stefan Bräse (IOC) for general support, Gary Davidson, Caitlin Howell and Markus Niermeyer for comments on the manuscript.

Funding: DOPAMINET Molecular Networks of Dopaminergic Neurons in Chordates. BOLD (Biology of Liver and Pancreatic Development and Disease) - Marie Curie Initial Training Network (238821 to A.H.S.); Dopaminet FP7 (Seventh Framework Program) (223744 to U.L.).

Conflict of Interest: none declared.

REFERENCES

Ansorge,W.J. (2009) Next-generation DNA sequencing techniques. *Nat. Biotechnol.*, **25**, 195–203.

Balakrishnan,I. *et al.* (2003) Looking up data in P2P systems. *Commun. ACM*, **46**, 43–48.

Ball,C.A. *et al.* (2004) Funding high-throughput data sharing. *Nat. Biotechnol.*, **22**, 1179–1183.

Berners-Lee,T. (2006) Linked Data. W3C Design Issues. *International Journal on Semantic Web and Information Systems*, Vol. 4, W3C, 1.

Birbeck,M. and Adida,B. (2008) *RDFa Primer. W3C Notes*. W3C Working Group Note 14 October 2008.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN*, **30**, 107–117.

Campbell,P. (2009) Data's shameful neglect. *Nature*, **461**, 145.

Falagas,M.E. *et al.* (2008) Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J.*, **22**, 338–342.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39** (Database issue), D800–D806.

Gstaiger,M. and Aebersold,R. (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.*, **10**, 617–627.

Jensen,L.J. and Bork,P. (2010) Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS Biol.*, **8**, e1000374.

Lagoze,C. *et al.* (2002) *The making of the Open Archives Initiative Protocol for Metadata Harvesting*. Library Hi Tech, Vol. 21, pp. 118–128.

Lewandowski,D. and Mayr,P. (2006) Exploring the academic invisible web. *Libr. Hi Tech*, **24**, 529–539.

Liebel,U. *et al.* (2004) 'Harvester': a fast meta search engine of human protein resources. *Bioinformatics*, **20**, 1962–1963.

McKiernan,G. (2005) E-profile: Scirus: For Scientific Information Only. *Libr. Hi Tech News*, **22**, 18–25.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Pepperkok,R. and Ellenberg,J. (2006) High-throughput fluorescence microscopy for systems biology. *Nat. Rev. Mol. Cell. Biol.*, **7**, 690–696.

Schadt,E.E. *et al.* (2010) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.*, **11**, 647–657.

Schuler,G.D. *et al.* (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

Swedlow,J.R. and Eliceiri,K.W. (2009) Open source bioimage informatics for cell biology. *Trends Cell. Biol.*, **19**, 656–660.

Szklarczyk,D. *et al.* (2010) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39** (Database issue), D561–D568.

Valentin,F. *et al.* (2010) Fast and efficient searching of biological data resources–using EB-eye. *Brief. Bioinform.*, **11**, 375–384.