SciNet: Augmenting Access to Scientific Information

Tuukka Ruotsalo[1], Kumaripaba Athukorala[2], Antti Ajanki[1], Mehmet Gönen[1], Murad Kamalov[1], Antti Oulasvirta[1], Chiwei Wang[1], Lilu Xu[2], Matti Nelimarkka[1], Samuli Hemminki[2], Sourav Bhattacharya[2], Petteri Nurmi[2], Dorota Glowacka[2], Patrik Floréen[2], Giulio Jacucci[2], Petri Myllymäki[2], Samuel Kaski[1,2]

[1] Helsinki Institute for Information Technology (HIIT), Aalto University, Espoo, Finland
[2] Helsinki Institute for Information Technology (HIIT), University of Helsinki, Finland
first.last@hiit.fi

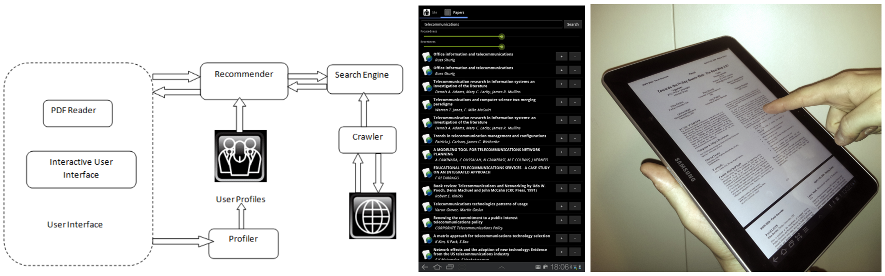
Abstract. Information needs of researchers are increasingly personalized, tailored to the state of knowledge of the user about different topics, dependent on the work context, and part of an interactive process, where users are engaged with the scientific information space. We aim at revolutionizing the way scientific information can be accessed. This vision is realized as the SciNet system that enables interactive scientific information access through personalized search and user profiling that are constructed by monitoring the users' behavior and allowing the users to interact with the underlying user models.

An amount of scientific product is estimated to be millions of publications worldwide per year; the growth rate of PubMed alone is now 1.8 paper per minute[3], and Google Scholar indexes 2.93 million articles for the year 2011[4]. Still we are only in the early days of a digital revolution, one which will have a deeper and more disruptive impact on scientific publishing and data distribution [1]. The problem of communication that the scientific community faces is shifting from retrieval of suitable materials to support every day work of researchers; to augment scientific work. In particular, scientific work has specific challenges that we tackle:

1. *Semantic gap.* The information needs of researchers are often exploratory, and drifting while users are interacting with the system. This can cause mismatch between the queries and the documents under interest.
2. *Cognitive gap.* The relevance of a document to the user's information need depends on the user's cognitive state. A novice user is likely to prefer authorative and well cited articles, while an expert focuses on recent work under narrower topics.
3. *Contextual gap.* The information needs of researchers are tailored to their work context. A scientists seek for very different type of information when writing an article, or participating in a meeting.

To address the above-mentioned challenges, we have constructed a prototype system SciNet illustrated in Figure1. It is an application framework consisting of three separate applications: search engine, user profiler, and recommender. The search engine and the recommender retrieve documents based on a query provided by the user,

[3] http://www.ncbi.nlm.nih.gov/pubmed/
[4] http://scholar.google.fi/



telecommunications

- Office information and telecommunications — *Russ Shurig*
- Office information and telecommunications — *Russ Shurig*
- Telecommunication research in information systems an investigation of the literature — *Dennis A. Adams, Mary C. Lacity, James R. Mullins*
- Telecommunications and computer science two merging paradigms — *Warren T. Jones, F. Mike McGuirt*
- Telecommunication research in information systems: an investigation of the literature — *Dennis A. Adams, Mary C. Lacity, James R. Mullins*
- Trends in telecommunication management and configurations — *Patricia J. Carlson, James C. Wetherbe*
- A MODELING TOOL FOR TELECOMMUNICATIONS NETWORK PLANNING — *A CAMINADA, C OUSSALAH, N GIAMBIASI, M F COLINAS, J KERNEIS*
- EDUCATIONAL TELECOMMUNICATIONS SERVICES - A CASE-STUDY ON AN INTEGRATED APPROACH — *F RI TARRAGO*
- Book review: Telecommunications and Networking by Udo W. Pooch, Denis Machuel and John McCahn (CRC Press, 1991) — *Robert E. Kinicki*
- Telecommunications technologies patterns of usage — *Varun Grover, Martin Goslar*
- Renewing the commitment to a public interest telecommunications policy — *CORPORATE Telecommunications Policy*
- A matrix approach for telecommunications technology selection — *K Kim, K Park, S Seo*
- Network effects and the adoption of new technology: Evidence from the US telecommunications industry — *S K Majumdar, S Venkataraman*

Fig. 1: Technical architecture of the SciNet system (left), the search user interface (middle) and the custom PDF reader (right) .

and rank the documents based on their topical likelihood. We utilize Latent Dirichlet Allocation to produce a semantic representation of documents in a latent topic space. We employ this powerful representation also as a user model and provide a learning to rank mechanism by inferring the most likely documents for each user in the retrieval phase. To reduce the cognitive gap, we harness several document measures including authorativity computed using citation networks, and recency based on the age of the document. The ranking is based on topics, but is dependent on users' preferences on these measures.

The preferences of the users are captured in a context-aware user model and estimated based on a variety of relevance feedback. The user profiler initially estimates weights for the topic model based on user's publication history crawled from external Web sources, such as Microsoft Academic Search [5], and the digital libraries of the ACM and the IEEE. Users can parametrize weights for the topics in a user interface or vote for documents in a search user interface. In addition, we have developed a custom PDF reader that tracks users reading behavior by capturing the the the user spends on reading the text on the screen.

The SciNet prototype is being deployed to be used by test users and we are investigating how it changes the way researchers search for scientific information. The system currently indexes over 20 million resources. Certain data included herein are derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011, the Digital Library of the Association of Computing Machinery (ACM), the Digital Library of Institute of Electrical and Electronics Engineers (IEEE), and the Digital Library of Springer.

References

1. Tim Berners-Lee and James Hendler. Scientific publishing on the 'semantic web'. *Nature*, pages 1023 – 1025, 2001. April 26th.

[5] academic.research.microsoft.com