

Dr. Woei-jyh (Adam) Lee
University of Maryland, College Park, USA

TITLE

Multipath Search on Large Open Data

ABSTRACT

With wired and wireless networks connecting the whole world in the 21st century, data were generated in unpredictable rates. On top of “Time is Money”, data can be turned into fortunes in possible various ways. Traditionally, scientists design and conduct experiments to generate data. Since not long ago, huge amount of data have been generated and waited for scientists to analyze. Hidden knowledge will be discovered then searched in the field of data science.

Biologists are used to search specific data sources to learn and gain published results. They search knowledge of human genes on some gene database, of human diseases on some disease database, of articles on some citations database. Three most commonly used data sources about human genetics are Gene, OMIM (Online Mendelian Inheritance in Man), and PubMed hosted by the NCBI (National Center for Biotechnology Information, United States National Library of Medicine, National Institutes of Health). However, such direct searches might lead to abundant records and search results. To filter or rank among search results remain challenging tasks.

We proposed an integrated search methodology using multiple search paths (i.e. multipaths). Many data are related therefore connected in different degrees via “links”. To search citations about a given term, instead of using the term to directly search on PubMed, biologists can search on the Gene database then follow links onto the PubMed. Such search steps form a search “path”. We experimented multipaths using various categories of terms among Gene, OMIM and PubMed.

Our results demonstrated benefits on multipath searches that we conducted. We analyzed the “all-way intersection” of multipath results and were able to narrow down more relevant results respected to a search term. Among the above three data sources, links from OMIM records to Gene records may be least redundant. Links from and to PubMed records may augment huge number of lower ranked data. Links between PubMed records and Gene records double the size of the traditional Gene and PubMed searches.

BIOGRAPHY

Dr. Woei-jyh (Adam) Lee received BSE degree from the National Taiwan University, MS degree from the Courant Institute at New York University, and PhD degree from the University of Maryland at College Park (UMD). He worked on distributed objects and fault tolerance at the AT&T Labs - Research in 1997. He focused on network software and management at the Bell Laboratories Research from 1998 to 2000. He visited the University of Southern California specializing in continuous media streaming and multimedia networking from 2002 to 2003.

He contributed in protein domain parsing and boundary prediction at the National Cancer Institute (NCI), National Institutes of Health (NIH) from 2004 to 2005. He was a fellow focusing on human genetics and genomics at the National Center for Biotechnology Information, National Library of Medicine, NIH from 2009 to 2012. He became a special volunteer working on computational modeling for cancer progression and metastatic at the NCI, NIH from 2012 to 2013. He was also affiliated with the Center for Bioinformatics and Computational Biology and the Institute for Advanced Computer Studies at UMD.

He is currently a faculty of Information Systems at the Robert H. Smith School of Business at UMD since 2012. His research interests include information integration, data analytics and mining, literature-based discovery, performance simulation and evaluation, bioinformatics and computational biology, human genomics and genetics, and cancer biology. He has two US Patents and is a member of the IAENG and the CAPA.