Exposing and Exploring Academic Expertise with Virtu

Luanne Freund
Univ.of British Columbia
Vancouver, Canada
luanne.freund@ubc.ca

Kristof Kessler
Univ.of British Columbia
Vancouver, Canada
kristof.kessler@alumni.ubc.ca

Michael Huggett
Univ.of British Columbia
Vancouver, Canada
m.huggett@ubc.ca

Edie Rasmussen
Univ.of British Columbia
Vancouver, Canada
edie.rasmussen@ubc.ca

ABSTRACT

The Virtu academic expertise finder system is presented. Virtu is built on the Apache Solr platform using data from Mendeley, a social network and bibliographic management system. Virtu takes a task-based approach to expertise, exposing and giving the user control over dimensions of expertise that are more or less desirable depending on the type of expert-finding task. The search interface supports information interaction and exploration through a number of browsing and filtering tools, including facets and sliders.

Author Keywords

Expertise finding; information retrieval; direct manipulation interface; task-based social network.

ACM Classification Keywords

H.3.3 Information search and retrieval
H.5.2 User Interfaces

General Terms

Human Factors; Design

INTRODUCTION

Finding individuals with relevant subject expertise is an important task in enterprise settings, as well as in broader domains such as academia, research and development, and human resources. One approach outside organizational settings is to mine the growing number of task-based social networks found on the Internet. Unlike relationship-based social networks such as Facebook, task-based networks such as Flickr (photography) and Mendeley (academic research and publication) are designed around shared activities and interests, and thus can be used as a source of data for identifying individuals and groups with particular skills and knowledge. User profiles and associated artifacts indicate the areas of expertise, and the connections within the network provide a means of identifying individuals with similar backgrounds and interests.

This paper presents Virtu, an expert-finding system in the academic domain that works on a dataset made available by Mendeley, a Web-based social network and bibliographic management tool. Virtu was custom-built for the HCIR challenge using the open-source Apache Solr search platform. Virtu takes a task-based approach to expertise, recognizing that different dimensions of skill and knowledge are more or less desirable depending on the searcher's work task. By identifying and exposing these dimensions and providing visualizations, facets and direct manipulation tools, Virtu supports focused retrieval as well as highly interactive and exploratory searching by the user.

RELATED SYSTEMS

There is a long history of research in expertise finding in fields such as Artificial Intelligence, Computer Supported Cooperative Work and Information Retrieval. In organizations, much of enterprise research and system development uses proprietary data drawn from corporate directories, intranets, and email. Such systems take different forms, including question routing and/or answering systems [1], matchmaking or team-building systems [2], and expert recommender systems [3].

Academic expertise systems are less common, but have become more prevalent due to the large amount of available networked data on academics and their publications. The AcademTech system [4] is one example of an academic expertise finder that draws on data from the Web. It is built on the Terrier IR platform, and claims to be the first expert search system to use a faceted search interface. The available facets, however, are quite limited: university, location, and number of publications. ArnetMiner is a Web-based academic expertise search system[1][5]. It offers searchers a range of task-based options, such as finding an expert, a conference or a reviewer. Although it has a highly visual and information-rich interface, ArnetMiner offers the user little opportunity to filter search results. The Rank feature[2] allows the user to produce lists of academics sorted according to eight expert statistics: H-Index, Total Citations, Uptrend, Activity, Longevity, Diversity, Sociability, and New Star. While these are similar to the expertise dimensions used in Virtu, the ArnetMiner rankings are separate from the Expert finder, and multiple measures cannot be combined to refine a search.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HCIR 2012, October 4-5, 2012, Cambridge, MA
Copyright held by authors.

[1] arnetminer.org
[2] arnetminer.org/person-rank

VIRTU

The design process for Virtu began with an analysis of several common expertise finding tasks within academia: hiring, assembling a conference program, and identifying a candidate to give expert testimony. Considering each, a number of dimensions of expertise emerged (Table 1). For example, it may be useful in a hiring scenario to identify candidates who are well connected, and whose future potential is suggested by a strong reputation early in their career. To find experts to provide legal testimony, subject knowledge and reputation are likely to be most important. The six dimensions we identified are subject knowledge, applied knowledge, years of experience, reputation, connectedness and multidisciplinary. The Mendeley dataset was mined to extract indicators for each of these dimensions (Table 1).

Virtu provides searchers with the means to filter and refine queries through direct manipulation of these six dimensions. Searchers can articulate very specific queries based on a set of a priori requirements, or explore the dataset of profiles interactively to make unexpected discoveries. The system also supports prototype queries that retrieve profiles that are similar or related to a single user profile of interest.

	Sources of Evidence
Subject Knowledge	Number of academic publications (of types Book, Book Section, Conference Proceedings, Encyclopedia Article, Journal Article, Thesis)
Applied Knowledge	Number of publications in all non-academic publication types.
Experience	Years elapsed since date of first publication.
Reputation	- Number of readers in same main discipline as author - Publication outlets, ranked by readers - Number of professorial versus non-professorial readers of the user's publications
Connectedness	- Number of contacts - Number of co-authors per publication - Number of groups joined - Total number of members in groups joined
Multidisciplinarity	- Number of disciplines - Number of Sub Disciplines - Percent of total readers from disciplines other than the author's

Table 1 – Expertise dimensions and sources of evidence

The System

The Virtu search engine is based upon Apache Solr, an open source enterprise search platform. Major features of Solr include full-text search, faceted search and filtering, field limits and weights, and database integration[3].

Of particular interest in the dataset provided by Mendeley was data related to what users had actually *done* as authors and editors. The provided *publications* file included basic data on the document's title, authors, year, venue, and tags, as well as the number, status, country, and area of expertise of the document's readers. On examining the public methods in the Mendeley REST API, we found that the document-details method provided significant additional data that could be useful for free-text and faceted search, including publication abstract, its topic areas, editor names, related special-interest groups, document identifiers, keywords, publisher names, and website URL.

To access this data, we first extracted all unique publication IDs from the *publications* file, then requested the document details of each ID using the REST API. Due to rate-limiting of 500 requests per hour in the Mendeley API, the run took 12 days to complete, which we compiled to a document-details file (*docDetails*) containing 142,052 records. We also extracted a list of authors per publication and compared their names to the names within the Mendeley profiles. If distinct matches were found, which was true for almost every publication, we updated the relations between Mendeley profiles and publications to include these authors. We also removed apparently incorrect profile-publication links.

Nearly one-third (45,086 or 32%) of the records provided no data (i.e. returned an empty array). To make the most of available data, we replaced each of these with records from the original *publications* file, translated from the *publications* schema into the *docDetails* schema. We then wrote a collection of parsers to extract database-friendly CSV tables suitable for incorporation into Solr. The parsers collated wanted fields, filtered out unwanted fields, and replaced (sometimes long) character strings with simple integer lookup-key IDs. We extracted further profile information from a sample of approximately 300,000 profiles and used this to enrich the user interface, particularly with images and titles if available.

The resulting data set was imported into a relational database to which Solr was instructed to send Select statements. The fields included in the query result are then integrated into the Solr schema, and search fields and Facet Fields and Ranges are defined.

Ranking

Standard IR approaches, such as tf-idf ranking, are not sufficient to identify top experts in a field, especially when the amount of text representing each expert is limited. We opted to boost profiles during indexing based on a

[3] http://lucene.apache.org/solr/features.html

combination of three measures, each a linear combination of components with weights that sum to 1:

- Knowledge, based on number of publications, weighted by the popularity of the publication venue (40%), years of experience (40%), and the number of disciplines in which the user has published (20%);
- Reputation, based on the number of readers weighted by the readers' academic rank (30%), the average number of co-authors per paper (30%), the number of contacts (20%), and the number of groups joined weighted by group size (20%). The user also received a 20% bonus to their score if they had served as editor of a publication;
- Profile Completeness, based on the amount of content listed in the user's research interests (40%) and in the user's biographical information section (60%).

The final expertise score for an individual user is the average of their Knowledge, Reputation and Profile Completeness scores.

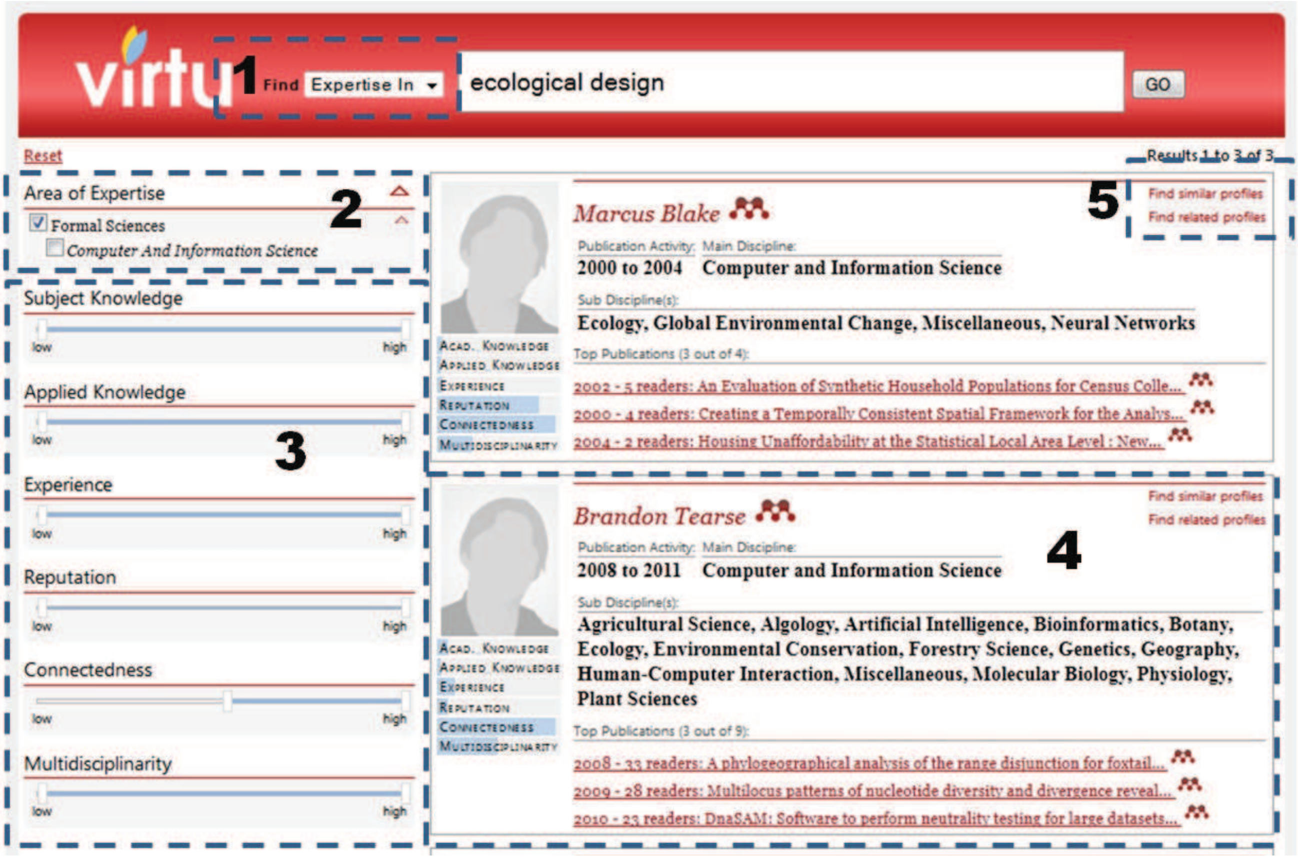
User Interface

We used the Solr Velocity Template Language in combination with JavaScript and Cascading Style Sheets to build the user interface. A screenshot of the Virtu search results page is shown in Figure 1. The components are -

Query input (1): A drop down box offers the option of searching for a person, for expertise in an area, or for all matches (default).

Discipline Facets (2): The main disciplines listed in profiles are presented as hierarchical facets, grouped into broad categories at the top level. Checkboxes allow the searcher to select one or more disciplines as limits on a keyword query, or to browse through profiles within a given discipline. The discipline facets are dynamically updated for each query and are combined as a Boolean OR search if multiple disciplines are selected.

Expertise Sliders (3): The six expertise dimensions are manipulated directly using dual-control sliders that set lower and upper bounds on each measure. The underlying values for each dimension are calculated as a linear combination of the measures in Table 1 and mapped to the range 0 to 100. The median value of each measure is set to the mid-point of the slider. Search results are updated as soon as a slider handle is released. Sliders are implemented using the jQuery UI Sliders library.



ecological design

☑ Formal Sciences
□ Computer And Information Science

Marcus Blake

Publication Activity: 2000 to 2004

Main Discipline: Computer and Information Science

Sub Discipline(s): Ecology, Global Environmental Change, Miscellaneous, Neural Networks

Top Publications (3 out of 4):

2002 - 5 readers: An Evaluation of Synthetic Household Populations for Census Colle...

2000 - 4 readers: Creating a Temporally Consistent Spatial Framework for the Analys...

2004 - 2 readers: Housing Unaffordability at the Statistical Local Area Level : New...

Brandon Tearse

Publication Activity: 2008 to 2011

Main Discipline: Computer and Information Science

Sub Discipline(s): Agricultural Science, Algology, Artificial Intelligence, Bioinformatics, Botany, Ecology, Environmental Conservation, Forestry Science, Genetics, Geography, Human-Computer Interaction, Miscellaneous, Molecular Biology, Physiology, Plant Sciences

Top Publications (3 out of 9):

2008 - 33 readers: A phylogeographical analysis of the range disjunction for foxtail...

2009 - 28 readers: Multilocus patterns of nucleotide diversity and divergence reveal...

2010 - 23 readers: DnaSAM: Software to perform neutrality testing for large datasets...

Figure 1: Virtu Search results page

Profile (4): Search results are displayed as index cards that include basic profile information such as name, publication activity, main discipline, sub disciplines derived from publications, the user's top three publications, and an image and title if available in the dataset. Links to the user's Mendeley page and the pages of the publications are also provided. Each profile shows the user's expertise score components as six horizontal bars representing the score for each dimension.

Find Similar and Related (5): Two links provided in each index card trigger prototype queries based on the selected profile of interest. *Find similar* retrieves profiles in the same discipline that have similar values on the expertise measures (+/-10 %), by adjusting the values of the Expertise Sliders. *Find related* uses data on co-authors, connections and groups to retrieve profiles that are most closely linked to the profile of interest.

DISCUSSION AND CONCLUSIONS

Virtu is a proof of concept and a work in progress. The main contribution of the system is the functionality of the expertise sliders, which allow the user to manipulate the results directly based on a set of abstract dimensions of expertise. A more granular approach was considered in which the user would have the ability to filter by discrete features, such as number of publications, number of groups, date of last publication etc., but we opted to give the user slightly less control and less transparency in favour of the improved usability of reducing these many features to six broad and well-defined dimensions.

The system is limited by the sparseness of the dataset—a perennial problem for expert finding systems that depend on profiles filled in manually by the users themselves. Given the choice early in the development process to rely only on the data available from Mendeley or to spend time harvesting additional data from the Web, we opted for the former in order to focus our efforts on implementing the design concept. We recognize that the effectiveness of the system is reduced as a result, but believe that the multiple modes of interaction will nevertheless allow for rich exploration within this dataset. External data will be used to supplement profiles in future iterations of the system.

The first priority for future work is conduct experiments to test and refine the field weights. With no training data available, ad hoc weights have been assigned in the current version of the system. User studies that assess the retrieval performance and the interface features are also needed to develop a more mature, usable system.

ACKNOWLEDGEMENTS

Thanks to Schuyler Lindberg for his valuable help with this project. This research was funded by the Next Generation Information Appliances Project (NGAIA) project of the GRAND Network of Centres of Excellence, Canada.

REFERENCES

1. Ackerman. M.S. Augmenting organizational memory: a field study of answer garden. *ACM Trans. Inf. Syst.* 16, 3 (1998), 203-224.

2. Foner, L.N. Yenta: a multi-agent, referral-based matchmaking system. In *Proc. of the first international conference on Autonomous agents* (AGENTS '97). ACM Press (1997), 301-307.

3. McDonald, D.W. and Ackerman, M.S. Expertise recommender: a flexible recommendation system and architecture. In CSCW '00: Proc. ACM Conference on Computer Supported Cooperative Work, ACM Press (2000), 231–240.

4. McDougall, D. & Macdonald, C. Expertise search in academia using facets. In *Proc.SIGIR conference on Research and development in information retrieval,* ACM Press (2009), 834-834.

5. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. ArnetMiner: extraction and mining of academic social networks. *Proc.SIGKDD international conference on Knowledge discovery and data mining,* ACM Press (2008), 990-998