"Idea Maps" of Research Fields

Jeff Alstott
Indiana University
jeffalstott@gmail.com

Bin Chen
Indiana University
binchen@indiana.edu

Diep Thi Hoang
Indiana University
dihoang@indiana.edu

ABSTRACT

Increasing digital distribution of scientific literature has lead to more effective literature based discovery and deeper analysis of research fields as complex networks. One such network is the "idea map" of a research field, in which the concepts of the field are linked to one another on the basis of how frequently they are written about together. These idea maps can be explored up close, to see the relationships between ideas in the field and how they change over time. These networks can all be studied on a systems level, in which the gross organization of ideas can vary greatly from field to field.

Categories and Subject Descriptors

H.1.1 [**Information Systems Models and Principles**]: General systems theory

General Terms

Documentation, Human Factors, Management, Measurement

Keywords

idea map, information science, history of science, network theory, literature-based discovery

1. INTRODUCTION

As information continues to become easier to access, systems for literature based discovery of research fields continues to expand in scope and quality [4]. Additionally, scientific analysis of scientific activity reaches ever increasing levels of detail and breadth, such as co-authorship network and citation network analysis [5, 1]. One of these many important elements of research fields is their structure of ideas, the way concepts are linked together [2]. The words used in a field and how they are frequently linked together form what we term "idea maps", which contain information not only on what ideas are commonly thought about together, but also as to the overall shape and structure of the field. The purpose of the present study was two-fold. First, to create a user friendly application for exploring idea maps. Second, to analyze the network theoretical properties of these idea maps, which we hypothesized would vary between research areas in significant ways.

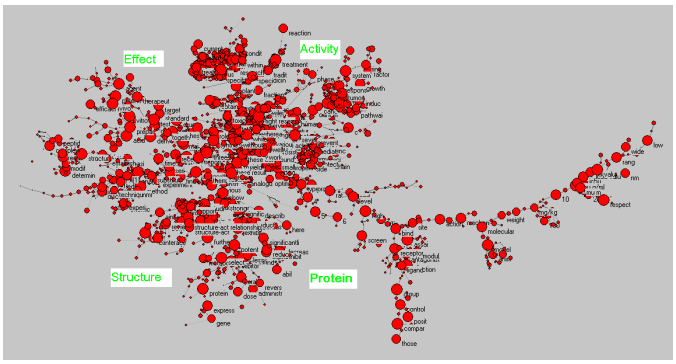

Figure 1: The idea map of Medical Chemistry in 2007. Only connections above a threshold strength are shown, save those necessary to make a minimum spanning tree.

2. DATA ACQUISITION AND PROCESSING

Nine research fields to be analyzed were identified from Thomson Scientific Journal Citation Report for 2007: Cell Biology, Chemistry (Inorganic), Chemisty (Organic), Chemistry (Medical), Computer Science, Ecology, Electrical Engineering, Geology, and Mathematics. The top 20 journals were selected for each field, on the basis of 5-year Impact Factor rank. We downloaded all abstracts and titles from 1991-2008 for each journal through ISI's Web of Knowledge.

All articles with empty abstracts were removed, and the remaing artile abstracts and titles went through several steps of processing: canonicalizing the abbreviations, lowering text case, removing common English stop words, eliminating punctuations, and stemming. Each processed abstract and title paper was then transformed into a bag of unigram terms (single words) and bigram terms (two consecutive words), with duplicates eliminated. The bags of terms were divided by field and year, and for each such group terms were thresholded; terms occuring in fewer than 10 abstracts or .5% of all abstracts, whichever was higher, were removed.

The bags of terms for each field-and-year group were then combined to create a binary abstracts x terms matrix $\mathbf{A}$ representing that group. Each abstract was a row, each term was a column, with a 1 indicating if a term was in a particular abstract and a 0 if not. The "term co-occurrence count" matrix $\mathbf{C}$ was created by multiplying $\mathbf{A}^T\mathbf{A}$. Dividing the elements of each row i of $\mathbf{C}$ by $\mathbf{C}_{i,i}$ produced the "co-occurrence strength" matrix. The values in this terms x terms matrix are the frequency that a given row term appeared in the same abstract as given column term. Each such matrix represents a weighted, directed network, such as that for Medical Chemistry in 2007 shown in Figure 1. These networks are "idea maps" of the research field for that year, which can be explored on a term by term basis or analyzed as an entire network.

3. USER APPLICATION

Envisioning the need of researchers with varying levels of computer skills to explore idea maps, we built the "Idea Maps Discoverer" Java desktop application to visualize the maps in various ways. Users select a research field to explore (ex. Cell Biology, Geology, etc.) and a year, and are presented with a list of terms present in the field at that time. Users then select a keyword, and the application displays a list of terms that co-occur with that keyword in abstracts. Co-occurrence strength with the keyword for each term in the list is displayed both as digits and a bar chart. Additionally, several network features of the keyword, such as centrality, clustering coefficient, etc., are displayed.

Idea Maps Discoverer can also generate an interactive backbone network map, similar to Figure 1, for a given field and year. Users can navigate around in this map and mouseover an individual term, which highlights its neighbors. Lastly, users can easily select two terms and and and produce a bar chart track the historical development of the weights between them.

These tools particularly aid literature-based discovery for users who may not yet know exactly what they are looking for. Simply looking at the initial list of co-occuring terms may reveal an unforeseen association. Wandering around the backbone map and interacting with its terms, users can visually identify communities of terms, which may indicate a topic in the field. Graphing the relationship between two terms over time can illuminate changes in the focus of a field, some examples of which are covered in section 4.

4. RESULTS AND DISCUSSION

With 9 fields, 18 years, and thousands of terms in each, there many, many possible relationships between individual terms to examine. A case study of one such relationship is covered in Figure 2, which shows the relationship of "gene" and "sequenc" (stemmed down from sequence, sequencing, etc.). The Human Genome Project started in 1990, and as such in the early 90s gene and sequence frequently occured in a same paper. Over time, the function of genes in a pathway drew more attention, and thus more papers studying genes focused on pathway, not sequence. Once the first human gene sequence was established in 1999, gene research diversified, including such topics as gene products and targeted drug discovery. In contrast, while talking about sequences,

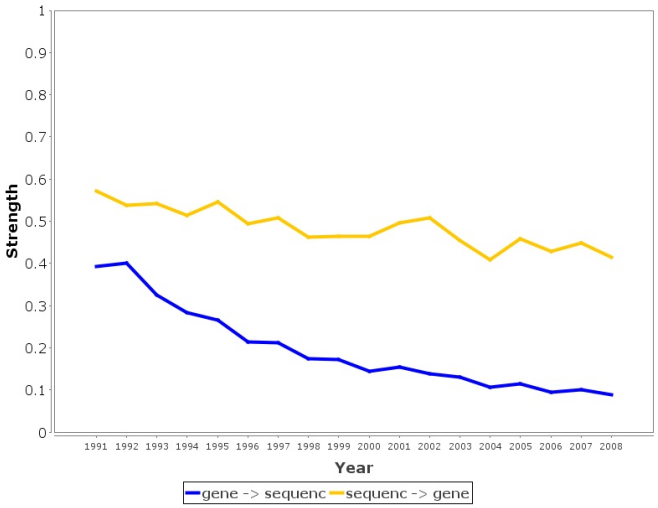

Figure 2: The co-occurrence strengths between the terms "gene" and "sequenc" in the Cell Biology field over the past 18 years.

people have continued to mention genes in the context of protein sequencing, though this trend is slowing [3].

In addition to the interactions of individual terms, we can consider the structural properties of entire fields, and particularly how they change over the years. Network measures such as path length, clustering coefficient, connection density, and small world index were calculated for each field each year. Over the nine field studied, the three chemistry fields, cell biology, and electrical engineering consisently stayed together over the 18 years sampled (Refer to Supplemental Information for figures). The other four fields took separate routes, such as decreasing connection density at various rates while the first five fields increased over the time period. However, while the first five fields seemed to "think alike" over the years, the other four did not; Mathematics, for example, had path lengths far larger than any other field.

Networks based on the simple co-occurrence of terms in a field's literature act as useful idea maps for the field, to be effectively used for internal exploration. They can also be a tool to compare and contrast the overall structures of separate research fields and identify schools of thought.

5. REFERENCES

[1] K. Borner, J. T. Maru, and R. L. Goldstone. The simultaneous evolution of author and paper networks. *Proc. Natl. Acad. Sci. USA*, 101(Suppl. 1):5266–5273, APR 6 2004.

[2] K. W. Boyack, K. Borner, and R. Klavans. Mapping the structure and evolution of chemistry research. *Scientometrics*, 79(1):45–60, APR 2009.

[3] B. Chen, 2009. Personal experience in the field.

[4] M. C. Ganiz, W. M. Pottenger, and C. D. Janneck. Recent advances in literature based discovery. *JASIST*, 2006.

[5] R. M. Shiffrin and K. Borner. Mapping knowledge domains. *Proc. Natl. Acad. Sci. USA*, 101(Suppl. 1):5183–5185, APR 6 2004.