Mapping the Information Science Domain
Ofer Arazy, Stan Ruecker, Omar Rodriguez, Alejandro Giacometti, Lu Zhang, and Su Chun

In the scholarly community, there has been for the last fifteen years an increasing interest in interdisciplinary projects, where researchers with different disciplinary backgrounds form teams to collaborate in addressing problems that are either not clearly within a single research domain, or that can best be approached by combining a variety of approaches. However, there also remain a number of research areas where interdisciplinary approaches seem to be warranted, but scholars in single disciplines have continued to work largely in isolation from their colleagues in other domains. We focus on one specific scientific field – information science – which is investigated in various disciplines: library and archival studies, computer science, management (and specifically, management information systems), and engineering, to name just a few.

Prior works in the area have discussed the disconnect between various areas within the broad field of information science, relying primarily on co-citation techniques and content analysis. These studies concluded that although all domains within this area are focused on the concepts of information, people, and (information and communication) technologies, the disciplines remain disjunct in terms of their disciplinary recognitions and topic of research. Despite some similarities in research methodology (“research approach” and “level of analysis”) (Glass et al., 2004), there are substantial differences in terms of co-citation patterns (Ellis et al., 1999; Sawyer & Huang, 2007), indicating the fields are distinct. Furthermore, although the research in these various areas may often appear to be similar on the surface, the contents of research are varied as well (Ellis et al., 1999). Glass et al. (2004) concluded that “overall, we see that there was minimal topic overlap” among the various disciplines, and Sawyer and Huang (2007) summarize that these areas of scholarship “have overlapping, but substantively different, research foci.”

Despite these interesting findings, extant research suffers from three primary limitations. First, previous studies are limited in scope, often focusing on just a subset of the academic units that publish articles in the domain. Second, prior research often relied on manual analysis, limiting the number of articles that are studied. Third, previous research in the area made little usage of information visualization techniques; recent advancements in the field of visualization have provided tools for exposing intricate relationships in data, and have the potential to reveal interesting patterns in interdisciplinary science.

Furthermore, although research on earlier data describes a disjoint domain, recent years have seen what appear to be the first signs of convergence in the various sub-fields within the broad information science discipline. We see more and more cross-departmental collaborations, journals with a much broader scope (e.g. JASIST), and there is a move across North America to amalgamate these various departments under one institutional roof (often referred to as the ‘information school’ or simply ‘iSchool’).

The objective of this paper is to address these previous limitations and map developments in the information science domain in the last fifteen years, using novel visualization techniques. We revisit the question of the divergence/convergence within the sub-disciplines, and propose a scalable method for analyzing (through both statistical analysis and advanced visualizations) the semantic and usage patterns in information science research articles. We focus our attention on three pre-defined topics that cross boundaries and are investigated in many sub-domains. For example the notion of ‘Trust’ was defined as “designing tools that would help alleviate the problem of trust in e-commerce, e.g. rating and reputation mechanisms”, and was represented using the following keywords: ‘trust’, ‘online’, ‘web’, ‘e-commerce’, ‘electronic commerce’, ‘e-business’, ‘trustworthiness’, ‘rating’, ‘reputation’, ‘reputation mechanism’, ‘system’, and ‘information system’.
Our method involved the following steps: (a) defining each of the three topic areas – ‘trust’, ‘text mining’, and ‘web services’ - through a series of search keywords; (b) running these keywords on the Thompson Reuters Web of Knowledge (http://wokinfo.com/) corpus of research articles, repeating the query separately for the 15 years under investigation, and choosing the 500 highest ranked (in terms of matching the query) research articles at each time period; (c) analyzing the relationships between the 500 articles in each time period using both co-citation and content analysis; and (d) developing novel visualization to expose how the patterns of relationships develop over the fifteen years.

Traditionally, citation patterns have been visualized using a network diagram, where the articles are represented as nodes and the citations are shown as lines connecting the nodes (e.g. Shen et al. 2006). Others have developed variations on this form of visual representation (Nakazono et al., 2006; Brandes & Willhalm, 2002; Ibekwe-SanJuan & SanJuan, 2009).

Our approach builds on the Eigenfactor method (Bergstrom & West, 2008), which applies an algorithm similar to Google’s PageRank (Brin & Page, 1998) to determine which nodes in a networks (in this case, journals) are most important, and how nodes connect with one another. The importance depends on where a journal resides in this mesh of citation links: the more citations a journal receives, especially from other well connected journals, the more important the journal is.

We model our visualization technique after the well-formed eigenfactor method (Stefaner, 2009) that utilizes Eigenfactor, and was used in the past to track the evolution of the Neuroscience domain (http://well-formed-data.net/archives/331/neuroscience-infoporn). That analysis pre-determined the association of journals to disciplines, and then analyzed the relationships between disciplines as they evolve over time. We, however, sought a more organic approach, which recognizes that a journal could publish articles belonging to various areas (and even a specific article may belong to more than one area). Thus, we employed the classification of articles to categories provided by the Web of Knowledge.

Building on prior work in the field (Ellis et al., 1999; Glass et al., 2004), we analyze relationships both in terms of content and citation patterns. For each of the three topics, and for each of the time periods, we analyzed the 500 research articles to create similarity matrices – in terms respectively of content and citation pattern – and then aggregated the data to calculate similarities between Web of Knowledge categories. These category similarity matrices were the input for our visualization method.

The results of our analysis are 3 pairs of visualizations, where each pair represents the evolution in the information science domain in terms of content and citation pattern for one topic. Figures 1a-c below depict the evolution in the ‘trust’ area, when using citation patterns as the method of analysis.

![Figure 1a: the evolution in the ‘Trust’ topic; each line represents a Web of Science category; each cluster is made up of similar categories, corresponding to a specific scientific discipline.](image-url)
Figure 1b: the “Information Science & Library Science” category, and how it evolved in terms of clustering with other categories.

Figure 1c: the cluster that includes the “Information Science & Library Science” category; on the right: the list of categories in this cluster; each colored line represents a category and its origins.

Figure 2, below, represents the visualization for the same topic – ‘Trust’ – this time based on the content of the research articles.
There are two interesting findings from our analysis. First, we see substantial differences between co-citation patterns and the patterns based on content (i.e. comparing Figures 1a and 2a). While the topical analysis shows one primary cluster, the co-citation visualization depicts many small clusters. Thus, while there is a substantial convergence in terms of contents, the information science field remains disjoint in terms of citation patterns. Some of the primary clusters correspond to the domains of management information systems, computer engineering, medical informatics, health information systems, physics, law, and computer science. Second, there are differences between topics. While the ‘Trust’ topic seem to have emerged from one area, and then broken into smaller sub-areas over the years (see Figure 1a), the ‘Text mining’ topic remains relatively consistent in terms of its decomposition to clusters.

We conclude that our novel data analysis and visualization methods have revealed patterns that have never been described before, highlighting the differences between topics of investigation within the broad information science discipline. In our next iteration of the project will include (a) the analysis of a larger corpus of research articles, trying to capture this entire domain, (b) refined data analysis methods, and (c) the exploration of new visualization techniques, which would help us to discover additional hidden patterns.

References


Acknowledgement
This research was supported in part by SSHRC, the Social Sciences and Humanities Research Council of Canada.