Research dynamics, impact, and dissemination: A topic-level analysis

Erjia Yan

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. Phone: (215)895-1459; Fax: (215)895-2494; Email: erjia.yan@drexel.edu

Abstract

In informetrics, journals have been used as a standard unit to analyze research impact, productivity, and scholarship. The increasing practice of interdisciplinary research challenges the effectiveness of journal-based assessments. The goal of this paper is to highlight topics as a valuable unit of analysis. A set of topic-based approaches is applied to a data set on library and information science publications. Results show that topic-based approaches are capable of revealing the research dynamics, impact, and dissemination of the selected data set. The paper also identifies a non-significant relationship between topic popularity and impact, and argues for the need to use both variables in describing topic characteristics. Additionally, a flow map illustrates critical topic-level knowledge dissemination channels.

Introduction

Informetric studies primarily focus on two threads of research. One employs journals as the unit of analysis: a paper is published in a journal and a journal is assigned to certain subject areas or research domains. The other thread is based on authorship: a paper is written by an author, an author is affiliated to an institution, and an institution is located in a certain geographical area. For the first thread, journals have been used as the de facto unit to address questions related to impact assessment (e.g., Pinski & Narin, 1976; Bollen, Rodriguez, & Van de Sompel, 2006), clustering and mapping (e.g., Ding, Chowdhury, & Foo, 2000; Van Eck & Waltman, 2010), and scientific trading (e.g., Borgman & Rice, 1992; Leydesdorff & Probst, 2009).

While journals serve as a fixed research instrument to study scholarship and design policies (e.g., Holton, 1978; Van Raan, 2004), science is becoming more collaborative and interdisciplinary (Metzger & Zare, 1999; Berners-Lee et al., 2006). Likewise, the effectiveness of journal-based analyses has been challenged (e.g., Rafols & Leydesdorff, 2009; Waltman & Van Eck, 2012). The prevalence of multidisciplinary, open access journals and online social media (e.g., Mendeley and CiteULike) has further intensified this issue (Evans & Reimer, 2009; Björk et al., 2010).

To reconcile this tension, scholars are calling for a more fine-grained analysis to classify research and assess impact. A set of solutions has been proposed: it includes the use of clustering techniques (e.g., Waltman & Van Eck, 2012; Boyack & Klavans, 2014) and probabilistic topic models (e.g., Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2007; Ramage et al., 2009) to identify research specialties or topics from papers. These studies focused largely on delineating topics and evaluating algorithmic performances. Consequently, we have a limited understanding of topic-level impact distribution and knowledge dissemination. A topic-level analysis enables research on science dynamics, influence assessment, and knowledge diffusion, filling the gap left by previous research.

1 Corresponding author
The central research question is how to use topic-based approaches to explore research dynamics, impact, and knowledge dissemination. Using a data set on library and information science (LIS) publications, the following questions will be addressed:

- What are the dynamic characteristics of topic popularity and impact in LIS?
- What topics are more self-contained? Do popular topics tend to be highly cited?
- What are the knowledge dissemination patterns of topics in LIS?

This study aims to expand the landscape of informetric research by exemplifying topics as a valuable unit of analysis. Although the findings are only applicable to LIS, the approaches and methods proposed in this paper are domain independent and should inform the studies of topic-level popularity and impact across different fields.

Literature review

Identifying research specialties through network methods

Bibliometric networks have been used to identify research specialties long before the recent proliferation of network analyses. Based on the level of analysis, bibliometric networks can be examined from paper-, author-, and journal-levels. The clustering of papers is made available through paper co-citation networks (Small, 1973) and paper bibliographic coupling networks (Kessler, 1963). The clustering results have been used to compare the topic similarity of papers (Kessler, 1963; Small, 1973) and to find intellectual turning points (Chen, 2004, 2006). Because co-citation relations reveal research specialties, scholars have used multidimensional scaling (e.g., White & McCain, 1998), factor analysis (e.g., White & Griffith, 1981), and pathfinder networks (e.g., White, 2003) to identify the subdivisions of research domains represented by author co-citation relations. At the journal-level, subject categories provide a fixed and consistent journal classification system, but the accuracy of this classification system has been questioned in recent years (e.g., Boyack, Klavans, Börner, 2005; Rafols & Leydesdorff, 2009). Accordingly, alternative journal classification schemes have been proposed through the use of journal-level bibliometric networks (e.g., Glänzel & Schubert, 2003; Leydesdorff & Vaughan, 2006; Zhang et al., 2010).

In addition to these co-occurrence-based approaches, there is a trend in bibliometrics to use hybrid networks to identify research specialties (e.g., Liu et al., 2010; Janssens, Glänzel, & De Moor, 2008; Boyack & Klavans, 2010; Zitt, Lelu, & Bassecoulard, 2011). It has been shown that hybrid approaches yielded more accurate clustering results than the use of only one type of network (e.g., Boyack & Klavans, 2010; Janssens, Glänzel, & De Moor, 2008). The limitation of these network-based analyses is that they rely on the interpretation of author or journal clusters. Thus, they may not provide sufficiently detailed information for a fine-grained analysis of the cognitive structure of research fields. This is largely attributed to the fact that authors or journals may represent a mixture of different specialties.

Identifying research specialties through topic models

Topic models have garnered much attention in recent years. We focus on two aspects: citation-aware topic models – models that consider cited references of text corpora, and dynamics topic models – models that are capable of distilling dynamic features from text corpora. Latent Dirichlet allocation (LDA)
models each document as a mixture of probabilistic topics. Through LDA, each topic is defined as a multinomial distribution over words (e.g., Blei, Ng, & Jordan, 2003; Rosen-Zvi et al., 2004; Steyvers, Smyth, & Griffiths, 2004; Blei & Lafferty, 2007; Tang et al., 2008; Tu et al., 2010). These topic models treat a document as a bag of words without considering its cited references, and can be referred to as citation-unaware models. Citation-aware models, on the other hand, used probabilistic latent semantic analysis (PLSA) or LDA as the generative process on both texts and citations (e.g., Cohn & Hoffman, 2000; Erosheva, Fienberg, & Lafferty, 2004; Dietz, Bickel, & Scheffer, 2007; Nallapati et al., 2008; He et al., 2009; Yang, Yin, & Davison, 2011; Masada & Takasu, 2012).

The dynamic aspect of topics has been investigated primarily through three approaches: post-hoc analysis (e.g., Griffiths & Steyvers, 2004; Hall, Jurafsky, & Manning, 2008), segmented approaches (e.g., Bolelli et al., 2009), and continuous-time models (Wang & McCallum, 2006). While post-hoc method used topic-document probability distributions to determine topic dynamics, segmented approaches divided document corpora into segments that have contingent time stamps. Both methods rely on the Markov assumption that the state at a single time point is independent from others. Wang and McCallum (2006) found that their non-Markov continuous-time model provided better prediction and more interpretable topical trends.

**Data and methods**

**Data set**

The field of library and information science (LIS) was chosen as a sample domain to address the proposed research questions. Publications in the Journal Citation Report subject category of Information Science and Library Science were harvested as the data set. The following filters were used: (1) year of publication: between 1955 and 2013; (2) document type: article, review article, or proceeding paper; (3) language: English; and (4) records that have no author, title, or cited reference were removed. The intermediate data set comprised 51,156 publications.

Because the LDA model needs meaningful words to conduct the inference, further data preprocessing of publications’ titles was implemented. To achieve this, (1) title words that have less than three letters (i.e., single-letter words and double-letter words) were removed from titles; (2) ten most frequently occurred words were removed from titles; and (3) those words that occurred in less than three publications were also removed; (4) finally, publications whose titles have less than three words were removed from the data set. The final resulted data set comprised 47,137 publications.

**The LDA model**

The Latent Dirichlet Allocation (LDA) model was proposed by Blei, Ng, and Jordan (2003) to identify topics from large publication corpora. The model assumes that words for each paper are derived from a mixture of topics and each topic follows a multinomial distribution and is defined on a collection of vocabulary. The basic procedures are (Wang & Blei, 2011, p. 450):

For each document in the data set:

- Draw topic proportions $\theta_d \sim \text{Dirichlet} (\alpha)$
- For each word $n$
  - Draw topic assignment $z_{d,n} \sim \text{Mult}(\theta_d)$
\( \theta_d \) is the per-document topic proportion for the \( d \)th document; \( z_{d,n} \) is the per-word topic assignment for the \( n \)th word in the \( d \)th document; \( w_{d,n} \) is the \( n \)th word in the \( d \)th document. To better understand LDA, we refer to the graphic model with plate notations (Blei, 2012, pp. 78, 81): \( N \) in Figure 1 denotes the number of words in a document, \( D \) denotes the number of documents, and \( K \) denotes the number of topics.

The graphic model shows the generative process of LDA. It assumes that each document comprises multiple topics in different proportions (\( \theta_d \sim \text{Dirichlet} (\alpha) \)); for instance, the current paper is mainly about topic modeling (high proportion), but also covers research dynamics (medium proportion), citation analysis (medium proportion), and knowledge diffusion (medium proportion). These topics are then derived from the per-document topic proportions (\( z_{d,n} \sim \text{Mult}(\theta_d) \)) and words in the document are drawn...
from these topics \((w_{d,n} \sim \text{Mult}(\beta_{z,d,n}))\). The generative process is an “imaginary random process by which the model assumes the documents arose (p. 78).” Words in a document are the only observed variable (highlighted in Figure 1); per-word topic assignment, per-document topic proportions, and topics are the latent variables to be identified. To achieve this, we rely on the inference process. This process uses the observed variable in the joint distribution to compute the conditional distributions (p. 80): \(p(\beta, \theta, z, w) = \prod_{i=1}^{K} p(\beta_i | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K, z_{d,n}})\). For extensive discussions on computing the posterior distribution from the joint distribution, we refer to Steyvers and Griffiths (2007), Blei and Lafferty (2009), and Blei (2012).

One advantage of the LDA model is that the words of each document are predicated on a mixture of topics, allowing a document to be assigned into multiple topics. This multi-assignment strategy delivers more flexible estimations than the classic co-citation analysis (e.g., White & McCain, 1998) or distance-based clustering (e.g., Yan, Ding, & Jacob, 2012). Limitations of the LDA model, however, arise from its rigid assumptions (Blei, 2012): the model assumes that the order of the words in the document is not relevant, nor does the order of documents in the data set, and the model also assumes that the number of topics is known and fixed. These assumptions help improve the computational efficiency, but they also cause complications when a text corpus does not comply with these assumptions. Solutions to relax these assumptions have been provided (see Blei, 2012). Previous studies on examining the cognitive structure of LIS have employed a variety of approaches and identified different numbers of research clusters, ranging from a few to more than a dozen (e.g., Åström, 2007; Järvelin & Vakkari, 2007; Prebor, 2010; Sugimoto et al., 2011; Yan, 2014; Zhao & Strotmann, 2014). Considering these scholarly efforts in relation to the size, diversity, and duration of the current data set, the number of topics was set at 50 for this study. This allows for a detailed diachronical analysis of the cognitive landscape of LIS. The result is evaluated in two ways: first, through Jensen-Shannon divergence (JSD); second, through a qualitative comparison of clusters obtained from previous studies (see discussion section).

**Topic popularity and impact**

We use Figure 2 to show the major steps to calculate topic popularity and impact. We started off on downloading scholarly data from academic databases. The data set was then prepared in two ways. One involved the use of citing articles and cited references to construct a paper citation network; the other took publication titles and used these textual data as the input for topic modeling. The raw input data were then processed based on certain pre-defined rules as to what words should be processed by a certain topic model. The Stanford Topic Modeling Toolbox (Stanford TMT: http://nlp.stanford.edu/downloads/tmt/tmt-0.4/) was used to perform the topic modeling. Other recognized toolkits include MALLET (http://mallet.cs.umass.edu/topics.php) and a series of open-source software packages by the Blei group at Princeton (https://www.cs.princeton.edu/~blei/topicmodeling.html). One output files from these topic model implementations is the document-topic proportion file (or simply the \(\theta\) file). By aggregating \(\theta\) for each topic, we are informed by the topic popularity; by multiplying \(\theta\) with paper citation network \((\theta' \times \text{PCN} \times \theta)\), we can obtain a topic citation network – this in turn informs us of the topic impact and topical knowledge paths. The way to calculate topic popularity and impact as well as topical knowledge paths is elaborated in the following paragraphs.
Figure 2. A flow chart showing the steps to obtain topic popularity and impact

Topic popularity was calculated through $\theta_d$, the per-document topic proportion for document $d$. It shows that if the topic proportion for a certain topic is high among a number of papers, this topic is considered as popular (e.g., Griffiths & Steyvers, 2004). Figure 3 shows an example of assigning three papers into five topics. As illustrated, the popularity of each topic $Pop(t(j))$ can be calculated through aggregating $\theta_{d,t(j)}$ for each topic $t(j)$.

![Table showing topic proportions for three papers](image)

The topic popularity score $Pop(t(j))$ for topic $j$ was calculated through aggregating $\theta_{d,t(j)}$:

$$Pop(t(j)) = \sum_d \theta_{d,t(j)}$$

where $t(j)$ denotes $j$th topic and $\theta_{d,t(j)}$ denotes the per-document proportion of document $d$ for topic $t(j)$. The topic popularity for topic $j$ in year $t$ can be expressed as:

$$Pop(t(j), t) = \sum_{d|py(d)=t} \theta_{d,t(j)}$$

where $py(d)$ denotes the publication year of document $d$.

For an effective comparison of topic popularity across different years, $Pop(t(j), t)$ was normalized by the sum of popularity scores of all topics for year $t$ (i.e., the total number of publications of year $t$):
\[ N_{\text{Pop}}(t(j), t) = \frac{\sum_d | \text{py}(d) = t \theta_{d,t(j)} \times c_r(d) = \text{py}(d) |}{\sum_j \sum_d | \text{py}(d) = t \theta_{d,t(j)} \times c_r(d) = \text{py}(d) |} \]

Using \( N_{\text{Pop}}(t(j), t) \) as the dependent variable and year of publication as the independent variable, a single factor linear regression model can be formed. Slopes of the modeled regression curve were used to measure the trend of topic popularity.

For each topic, the average year of publication informs us of its development stages. It was calculated by multiplying \( \theta_{d,t(j)} \) with \( \text{py}(d) \), divided by the aggregated sum of \( \theta_{d,t(j)} \). The average year of publication of topic \( j \) can thus be calculated through:

\[ \text{py}(t(j)) = \frac{\sum_d (\theta_{d,t(j)} \times \text{py}(d))}{\sum_d \theta_{d,t(j)}} \]

Slopes of topic popularity can be further evaluated through \( z \)-score (Yan, 2014), from which labels of declining, fluctuating, and raising topics are assigned. The focus of this paper lies beyond the assignment of topics but on the examination of the relationship between topic popularity and impact. In the following paragraphs, we introduce a way to add citations to topics and form a topic citation network.

An internal paper citation network (PCN) was constructed. This citation network comprised 47,137 citing papers and their associated citation relations. The total number of internal citations was 93,932. It is the number of times papers in the network have been cited by other papers in the network. PCN will afford the investigation of the proposed research questions by the analysis of citation impact and knowledge dissemination within LIS. The choice of internal citations is consistent with other network-based citation analysis (e.g., Leydesdorff, 2007; Ma, Guan, Zhao, 2008; Yan & Sugimoto, 2011). These citation instances were then aggregated into the identified 50 topics based on \( \theta_d \). The topic citation impact \( CI(t(j), t) \) for topic \( j \) at year \( t \) is:

\[ CI(t(j), t) = \sum_{d | \text{py}(d) = t} \theta_{d,t(j)} \times c_r(d) \]

where \( CI(t(j), t) \) denotes citation impact of topic \( j \) at year \( t \), \( c_r(d) \) denotes the number of citations document \( d \) received. This fractional counting aligns with the citation impact assessment for authors (e.g., Egghe, 2008), journals (e.g., Zitt & Small, 2008), and institutions (e.g., Leydesdorff & Shin, 2011). For an effective comparison of topic citation impact across different years, \( CI(t(j), t) \) was normalized by the sum of citations of all topics for year \( t \) (i.e., the total number of citations received by papers published in year \( t \)):

\[ N_{\text{CI}}(t(j), t) = \frac{\sum_d | \text{py}(d) = t \theta_{d,t(j)} \times c_r(d) |}{\sum_j \sum_d | \text{py}(d) = t \theta_{d,t(j)} \times c_r(d) |} \]

Using the \( N_{\text{CI}}(t(j), t) \) as the dependent variable and year of publication as the independent variable, a single factor linear regression model can be formed. The slope of the modeled regression curve was used to measure the trend of topic impact.
Let PCN be the 47,137 by 47,137 paper citation network; $\theta$ is the per-document topic proportion for all documents that has a dimension of 47,137 by 50. The topic-to-topic citation network (TCN) can thus be obtained by matrix manipulation $\theta' \times PCN \times \theta$. The dimension of TCN is 50 by 50. Using the same example topic proportion in Figure 3 and a three-paper PCN, Figure 4 shows the way to form a topic citation network.

As with disciplines, topics may exhibit different citation behaviors: some topics may be more permeable and “inter-topical” whereas others may be more self-contained. Topic permeability was measured through topic self-citations:

$$Per(t(j)) = \frac{TCN_{jj}}{\sum_i TCN_{ij}}$$

where $Per(t(j))$ denotes permeability of topic $t(j)$, $TCN_{jj}$ denotes self-citation of topic $t(j)$, and $TCN_{ij}$ denotes citation from topic $t(i)$ to topic $t(j)$.

Currently, clustering and mapping techniques are largely designed for co-occurrence based networks (e.g., White & McCain, 1998; White, 2003; Rafols, Porter, & Leydesdorff, 2010). A different mapping strategy is needed to effectively represent a dense, small-scaled, directed network. In this study, shortest path was employed as the instrument to map knowledge flow among topics. Shortest paths characterize the most important knowledge paths in the topic citation network. Shortest path is typically defined in a distance-based network. The topic-to-topic citation network was thus transformed into a distance-based network through $\text{reverse flow width}_{i \rightarrow j} = \frac{1}{\text{number of citations from } j \text{ to } i}$ It shows that the more citations from one topic to another, the wider the knowledge flow (Yan, in press). All 2,500 paths (i.e., 50 *50 by including paths to themselves) were identified among the 50 topics through the Dijkstra algorithm. The most frequently occurred one-hop paths (i.e., paths that only involve two nodes) were extracted (see an example in Figure 5). They were used as the backbone of topical knowledge diffusion.
Results

Topics in library and information science

The 50 topics were labeled by the top five words that have the highest associations with each topic, as seen in Table 1. For each topic, average year of publication, peak year, popularity and slope, and citation impact and slope are also included. Because earlier years may not provide sufficient data points for the linear regression model, slopes for topic popularity were calculated based on publications between 1964 and 2013. Meanwhile, because papers published in 2013 were cited very sparsely due to a narrow citation window, slopes for topic impact were calculated based on publications between 1964 and 2012.

Table 1. Features of the 50 topics in library and information science

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
<th>Average year of publication</th>
<th>Peak year</th>
<th>Popularity</th>
<th>Slope (1964-2013)</th>
<th>Citation impact</th>
<th>Slope (1964-2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>factors-user-technology-use-effects</td>
<td>2001.69</td>
<td>2011</td>
<td>1103.09</td>
<td>5.19E-04</td>
<td>5261.53</td>
<td>1.06E-03</td>
</tr>
<tr>
<td>t2</td>
<td>papers-hot-biology-human-cell</td>
<td>1998.14</td>
<td>1993</td>
<td>613.08</td>
<td>2.54E-04</td>
<td>559.89</td>
<td>1.02E-04</td>
</tr>
<tr>
<td>t3</td>
<td>science-review-technology-issues-international</td>
<td>1992.80</td>
<td>1969</td>
<td>901.94</td>
<td>-8.10E-04</td>
<td>2144.57</td>
<td>-1.71E-04</td>
</tr>
<tr>
<td>t4</td>
<td>scientific-science-collaboration-international-bibliometric</td>
<td>2000.54</td>
<td>2013</td>
<td>1296.55</td>
<td>5.00E-04</td>
<td>4812.41</td>
<td>7.28E-04</td>
</tr>
<tr>
<td>t5</td>
<td>document-review-literature-supply-delivery</td>
<td>1996.89</td>
<td>1987</td>
<td>678.61</td>
<td>-2.28E-05</td>
<td>955.30</td>
<td>1.23E-04</td>
</tr>
<tr>
<td>t7</td>
<td>law-model-function-data-new</td>
<td>1997.83</td>
<td>2013</td>
<td>797.12</td>
<td>1.20E-05</td>
<td>2670.05</td>
<td>-1.65E-03</td>
</tr>
<tr>
<td>t8</td>
<td>data-using.spatial-gis-model</td>
<td>2003.21</td>
<td>2000</td>
<td>1018.78</td>
<td>5.37E-04</td>
<td>1578.77</td>
<td>5.45E-04</td>
</tr>
<tr>
<td>t10</td>
<td>model-theory-models-seeking-design</td>
<td>2000.84</td>
<td>2007</td>
<td>916.93</td>
<td>3.27E-04</td>
<td>2500.87</td>
<td>6.57E-04</td>
</tr>
<tr>
<td>t11</td>
<td>systems-support-decision-system-design</td>
<td>1995.48</td>
<td>1987</td>
<td>998.52</td>
<td>-5.11E-05</td>
<td>1415.88</td>
<td>9.19E-05</td>
</tr>
<tr>
<td>t12</td>
<td>collections-collection-development-special-problems</td>
<td>1991.39</td>
<td>1965</td>
<td>893.09</td>
<td>-1.11E-03</td>
<td>991.43</td>
<td>-6.51E-04</td>
</tr>
<tr>
<td>t14</td>
<td>cataloging-collection-book-circulation-hallmark</td>
<td>1986.85</td>
<td>1965</td>
<td>967.06</td>
<td>-1.89E-03</td>
<td>1030.31</td>
<td>-7.27E-04</td>
</tr>
<tr>
<td>t15</td>
<td>bibliographic</td>
<td>1997.69</td>
<td>1965</td>
<td>789.31</td>
<td>-8.15E-05</td>
<td>1553.20</td>
<td>1.16E-04</td>
</tr>
<tr>
<td>t17</td>
<td>books-book-literature-work-years</td>
<td>1984.75</td>
<td>1966</td>
<td>1002.28</td>
<td>-1.73E-03</td>
<td>721.69</td>
<td>-3.64E-04</td>
</tr>
<tr>
<td>t19</td>
<td>theory-making-communication-systems-practice</td>
<td>2001.15</td>
<td>2013</td>
<td>877.85</td>
<td>3.67E-04</td>
<td>1758.16</td>
<td>3.00E-04</td>
</tr>
<tr>
<td>t20</td>
<td>health-informatics-medical-care-sciences</td>
<td>2001.88</td>
<td>2002</td>
<td>900.54</td>
<td>2.81E-04</td>
<td>1182.73</td>
<td>3.47E-04</td>
</tr>
<tr>
<td>t21</td>
<td>law-legal-american-history-public</td>
<td>1988.46</td>
<td>1968</td>
<td>892.53</td>
<td>-1.52E-03</td>
<td>725.72</td>
<td>-1.00E-03</td>
</tr>
<tr>
<td>t22</td>
<td>classification-indexing-using-text-automatic</td>
<td>1995.68</td>
<td>1971</td>
<td>959.63</td>
<td>-4.07E-04</td>
<td>1140.08</td>
<td>-1.66E-03</td>
</tr>
<tr>
<td>t23</td>
<td>retrieval-using-query-relevance-document</td>
<td>1998.09</td>
<td>1971</td>
<td>1241.10</td>
<td>2.31E-05</td>
<td>1221.65</td>
<td>-1.50E-03</td>
</tr>
<tr>
<td>t24</td>
<td>empirical-online-technology-electronic-model</td>
<td>2004.23</td>
<td>2013</td>
<td>1143.17</td>
<td>7.92E-04</td>
<td>4033.81</td>
<td>1.38E-03</td>
</tr>
<tr>
<td>t26</td>
<td>health-cancer-communication-internet-among</td>
<td>2004.00</td>
<td>2006</td>
<td>976.63</td>
<td>3.94E-04</td>
<td>1318.15</td>
<td>5.05E-04</td>
</tr>
<tr>
<td>t27</td>
<td>language-medical-using-knowledge-data</td>
<td>2000.78</td>
<td>1998</td>
<td>1060.60</td>
<td>3.81E-04</td>
<td>776.74</td>
<td>1.00E-05</td>
</tr>
<tr>
<td>t28</td>
<td>use-university-electronic-students-resources</td>
<td>2000.05</td>
<td>1968</td>
<td>903.33</td>
<td>1.10E-04</td>
<td>2796.27</td>
<td>-6.85E-04</td>
</tr>
<tr>
<td>t29</td>
<td>scientists-science-researchers-new-with</td>
<td>1995.82</td>
<td>1990</td>
<td>862.59</td>
<td>9.03E-05</td>
<td>1757.27</td>
<td>1.18E-04</td>
</tr>
<tr>
<td>t30</td>
<td>more-century-better-people-work</td>
<td>1996.97</td>
<td>1965</td>
<td>684.85</td>
<td>6.16E-06</td>
<td>959.11</td>
<td>7.82E-05</td>
</tr>
<tr>
<td>t31</td>
<td>copyright-privacy-policy-issues-security</td>
<td>1997.55</td>
<td>1968</td>
<td>806.68</td>
<td>1.84E-05</td>
<td>895.64</td>
<td>2.11E-04</td>
</tr>
<tr>
<td>t34</td>
<td>impact-journals-citation-journal-scientific</td>
<td>2000.90</td>
<td>2013</td>
<td>971.52</td>
<td>3.94E-04</td>
<td>3540.85</td>
<td>6.46E-04</td>
</tr>
<tr>
<td>t36</td>
<td>learning-literacy-students-education-instruction</td>
<td>2001.11</td>
<td>2011</td>
<td>1017.50</td>
<td>2.66E-04</td>
<td>1953.58</td>
<td>5.42E-04</td>
</tr>
<tr>
<td>t37</td>
<td>management-systems-information-systems-issues-implementation</td>
<td>1998.38</td>
<td>1994</td>
<td>963.11</td>
<td>1.65E-04</td>
<td>2499.09</td>
<td>1.81E-04</td>
</tr>
<tr>
<td>t38</td>
<td>literature-citation-science-scientific-networks</td>
<td>1998.75</td>
<td>1974</td>
<td>1019.32</td>
<td>7.39E-05</td>
<td>3733.02</td>
<td>-1.83E-03</td>
</tr>
<tr>
<td>t40</td>
<td>telecommunications-mobile-market-policy-competition</td>
<td>1999.70</td>
<td>1992</td>
<td>1115.86</td>
<td>4.65E-04</td>
<td>1712.36</td>
<td>4.03E-04</td>
</tr>
<tr>
<td>t41</td>
<td>search-web-searching-user-users</td>
<td>2001.55</td>
<td>2007</td>
<td>888.70</td>
<td>4.42E-04</td>
<td>1870.88</td>
<td>-1.25E-04</td>
</tr>
<tr>
<td>t42</td>
<td>social-online-network-communities-networks</td>
<td>2004.77</td>
<td>2012</td>
<td>907.75</td>
<td>5.36E-04</td>
<td>1273.15</td>
<td>4.68E-04</td>
</tr>
<tr>
<td>t43</td>
<td>years-science-next-new-society</td>
<td>1996.60</td>
<td>1987</td>
<td>789.17</td>
<td>-1.17E-04</td>
<td>1020.69</td>
<td>1.58E-04</td>
</tr>
<tr>
<td>t44</td>
<td>business-technology-value-process-small</td>
<td>1999.95</td>
<td>1995</td>
<td>884.66</td>
<td>3.41E-04</td>
<td>2405.64</td>
<td>5.72E-04</td>
</tr>
<tr>
<td>t45</td>
<td>know-good-question-online-find</td>
<td>1997.54</td>
<td>1965</td>
<td>1044.31</td>
<td>1.31E-04</td>
<td>1955.48</td>
<td>2.22E-04</td>
</tr>
<tr>
<td>t46</td>
<td>new-education-future-librarianship-technology</td>
<td>1991.34</td>
<td>1971</td>
<td>932.57</td>
<td>-9.66E-04</td>
<td>1236.03</td>
<td>-2.40E-04</td>
</tr>
<tr>
<td>t47</td>
<td>data-digital-system-database-metadata</td>
<td>2000.83</td>
<td>1997</td>
<td>1066.95</td>
<td>4.53E-04</td>
<td>1206.70</td>
<td>7.64E-05</td>
</tr>
<tr>
<td>t49</td>
<td>digital-developing-africa-south-countries</td>
<td>2002.60</td>
<td>2011</td>
<td>878.81</td>
<td>3.62E-04</td>
<td>1248.06</td>
<td>4.69E-04</td>
</tr>
</tbody>
</table>
Based on average year of publication, 23 topics belong to the 2000s, 25 belong to the 1990s, and 3 belong to the 1980s. The three most recent topics are t42 on online social networks, t24 on online technologies, and t26 on health communications. The three most dated topics are t17 on books and literature, t14 on cataloging, and t21 on history of legal services. Popularity of the 50 topics ranges from 604 to 1,296 with a mean of 942. The three most popular topics are t4 on scientific collaboration, t23 on information retrieval, and t24 on online technologies. Measured by slope of popularity, 34 topics have positive slopes and thus have gained popularity; in the meantime, 16 topics have negative slopes and are thus becoming somewhat less visible in LIS. Citation impact of the 50 topics ranges from 560 to 5,262 with a mean of 1,879. Top three topics that have the highest topic impact are t1 on user and technology, t4 on scientific collaboration, and t16 on citation analysis. Measured by slope of citation impact, 32 topics have positive slopes and 18 topics have negative slopes.

**Topic popularity**

The dynamic aspect of topic popularity is assessed through $N_{Pop}(t(j), t)$. In Figure 6, x-axis denotes year of publication (1964-2013) and y-axis denotes $N_{Pop}(t(j), t)$ for top five topics that have the steepest popularity loss (left panel) and popularity gain (right panel), measured by slope.

![Figure 6](image_url)

**Figure 6.** Topics that have the steepest popularity loss (left panel) and popularity gain (right panel)

In Figure 7, while topics on cataloging, collection, and librarianship have experience the most noticeable popularity decrease, topics on online technologies, data and knowledge management, and health communications have the most evident popularity gain.

**Topic impact**
The dynamic aspect of topic impact is measured through $N_{CI}(t(j), t)$. In Figure 7, x-axis denotes year of publication (1964-2012) and y-axis denotes $N_{CI}(t(j), t)$ for top five topics that have the steepest impact loss (left panel) and impact gain (right panel), measured by slope.

Figure 7. Topics that have the steepest impact loss (left panel) and impact gain (right panel)

Compared with dynamics of topic popularity, the dynamics of topic impact is subject to higher degrees of variation. This is attributed to the fact that preeminent papers can receive up to several hundred citations, causing yearly topic impact fluctuations. Topics on literature, classification, and document retrieval were highly cited in the 1960s and 1970s; however, the concentration was shifted to online technologies, knowledge management, and journal citation impact analysis thereafter.

**Discussion**

**Statistical and qualitative evaluations of the results**

Jensen-Shannon divergence (JSD) was employed to evaluate the quality of the 50 topics. JSD is calculated based on word topic assignment. The lower bound of JSD is 0, the situation that two word topic assignments are identical; the upper bound is 0.7 ($\ln(2)$), the situation that two topics are completely different. Figure 8 shows the heatmap of the topic dissimilarity matrix measured by JSD.
Figure 8. A heatmap presentation of Jensen-Shannon divergence for the 50 topics

Figure 8 illustrates that most topics have high JSD scores with respect to other topics. This is a good sign that the topic model has successfully identified distinctive topics. We also see that some topics are topically related to others. This is expected because different topics may share a (small) collection of common lexicons (Blei & Lafferty, 2007); this is especially true for topics from a single field such as LIS. While the use of JSD is able to establish some statistical certainty, the real-world implications should also be examined. Thus, we compare the results of this study with previous analyses of the cognitive structure of LIS and use this as an opportunity to evaluate its empirical implications.

One category of analysis has primarily focused on the identification of the epistemological divisions of LIS. It has been argued that information science and library science have philosophical and theoretical differences (e.g., Brooke, 1980; Hjørland, 2000; Ma, 2012). Quantitative studies have supported such argument; for instance, through co-citation and bibliographic coupling analysis, studies have identified subfields in LIS: library science, information science and technology (information retrieval), and informetrics (Åström, 2002; Milojević et al., 2011; Waltman, Yan, Van Eck, 2011). Results obtained from this study align with the previous analyses in that the 50 topics can be grouped into these three subfields.

In addition top-level divisions, studies have also identified research clusters in LIS. The pioneering author co-citation analysis by White and Griffith (1981) used factor analysis to identify five clusters between 1972 and 1975, including scientific communication, bibliometrics, generalists, information retrieval, and precursors of LIS. The author co-citation analysis was later highlighted by White and McCain (1998) who
identified 12 research clusters between 1972 and 1995, including experimental retrieval, citation analysis, online retrieval, bibliometrics, general library systems, science communication, user theory, OPACs, imported ideas, indexing theory, citation theory, and communication theory. In a follow-up article by White (2003), a novel pathfinder network was used to visualize author co-citation relations and the same clusters were revealed (with the exception that citation theory was no longer a standalone cluster). A recent author co-citation analysis by Zhao and Strotmann (2014) has suggested a different set of clusters between 2006 and 2010. These include information behavior, bibliometric distributions, mapping of science, relevance, IR systems, webometrics, bibliometrics and science & innovation systems, use of e-resources, IS theory & foundation, knowledge management, and text categorization. Results obtained from this study are consistent with these previous efforts. For instance, it has delineated topics broadly related to informetrics (t4, t16, and t34), information retrieval (t23, t25, and t41), communication theory (t19), knowledge management (t32 and t37), indexing theory (t22), user theory (t1), information behavior (t10), and use of e-resources (t28). In the meantime, we also identified topics that were not included by previous studies, such as topics on health communication (26), health informatics (t26), electronic clinical record (t33), social networks (t42), software development (t6), and information systems (t11). This study has also identified precursory topics of some of these topics (see section on knowledge flow and inheritance).

By matching the results of this study with previous diachronical analysis, some consensus can be made. For raising trends, this study has found that topics on users, bibliometrics, online technologies, communications, health informatics, and knowledge management are becoming more popular in LIS. It confirms Larivière, Sugimoto, and Cronin’s (2012) study that words such as citation, impact, bibliometrics, use, management, health, clinical, and network have become popular title words in LIS between 1900 and 2010. A study by Milojević and colleagues (2011) has also found that title terms such as citation, impact factor, and web have a rising usage between 1989 and 2008. For declining trends, this study has shown that topics on collections, books and literature, indexing, and librarianship are becoming less popular. The result is consistent with Larivière, Sugimoto, and Cronin’s (2012) study.

The set of approaches used in this study has three advantages. First, the topic modeling technique works directly with words and does not require domain knowledge to interpret the clusters of papers, authors, or journals. From this aspect, this technique relates to co-word analysis (e.g., Janssens, Glänzel, & De Moor, 2008; Milojević et al., 2011; Yan, Ding, & Jacob, 2012). Different from co-word analysis, though, is that the topic modeling technique assigns words to clusters based on certain probabilistic distributions. Co-word analysis, on the other hand, typically partitions words to mutually exclusive clusters. Second, the number of topics identified in this study is considerably larger than those obtained from co-occurrence-based network analyses. The sizable number of topics grants a more granular analysis of the cognitive structure of LIS. In addition to probing into the top-level divisions or major clusters of LIS, this study is able to depict a more detailed research landscape. Such level of analysis is usually not attainable through co-occurrence-based network analysis due to the complexity of some densely connected co-occurrence clusters. Finally, the proposed dynamic analysis method has streamlined the diachronical analysis of topic popularity and impact. The rendering of yearly topic popularity is an improvement over time-sliced methods. Furthermore, the integration of citation information with topic analysis has created an opportunity to examine the inter-topic knowledge diffusion patterns in LIS. These advances have complemented co-occurrence-based network studies.
Journal-level topical diversity

To quantitatively measure the journal-level topical diversity, we use a heatmap (Figure 9) to show the distribution of topics on the most productive 30 journals in LIS. Cell values in the heatmap are calculated through aggregating $\theta_d$ for each journal under each topic and then normalizing by the total number of topic proportions for each journal. The journal-level topical diversity of journal $j(s)$ for topic $j$ can thus be expressed as $JTD_{j(s)} = \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)}$. Cell values range from 0 to 1 where 0 being no specialization in the topic and 1 being exclusively specialized in the topic.

\begin{align*}
J_{TD,j(s)}(c) &= \sum d \in j(s) \theta_d, t(j) \\
&= \sum_j \sum_{d \in j(s)} \theta_d, t(j) \\
&= \sum_j \sum_{d \in j(s)} \theta_d, t(j) \\
&= \sum_j \sum_{d \in j(s)} \theta_d, t(j)
\end{align*}

Figure 9. Heatmap presentation of journal topical specializations

While a handful of journals (e.g., INT J GEOGR INF SCI, J HEALTH COMMUN, J AM MED INFORM ASSN, and TELECOMMUN POLICY) have more specialized foci, the majority of journals have wider research specialties. The result suggests that these journals do not necessarily publish papers solely on one research specialty but contribute to an array of specialties. These journals may be able to represent LIS as a field of study but they may not have the granularity to signify the different research specialties within LIS.

Shannon entropy (Figure 10) is adopted to illustrate the dynamic changes of journal topical diversity. It measures, for each journal, the proportions of its papers under each topic:

\begin{align*}
H &= -\sum_{j=1}^{50} JTD_{j(s)} \ln JTD_{j(s)} = -\sum_{j=1}^{50} \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)} \ln \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)} \\
&= -\sum_{j=1}^{50} \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)} \ln \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)} \\
&= -\sum_{j=1}^{50} \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)} \ln \frac{\sum_{d \in j(s)} \theta_d, t(j)}{\sum_j \sum_{d \in j(s)} \theta_d, t(j)}
\end{align*}

Shannon entropy is a measure of dissimilarity in that the higher the value, the more diverse the content; different from JSD, though, it measures the congruence of a single object based on its variables whereas JSD is typically applied to measure the dissimilarity of two objects.
Figure 10. Shannon entropy for the most productive 30 journals in LIS

Diachronically, most journals in Figure 10 have either a steady or increased level of diversity, with the exception of two health informatics journals (J AM MED INFORM ASSN and J MED LIBR ASSOC) and two professional journals (LIBR J and ONLINE). A few journals have a steeper diversity gain, such as GOV INFORM Q, J DOC, and LIBR INFORM SCI RES. It indicates that these journals are embracing a broader research scope and publishing papers that cover more extensive topics. As journals are becoming more inter-topical and even interdisciplinary, simply relying on journal-level analyses to examine the substrate of various research fields may be less effective. Based on this evidence, therefore, we argue that topic-level studies are needed to propel the analysis to a newer level of granularity.

**Topic popularity vs. impact**

Similar to authors and journals, topics also possess multiple attributes. In addition to social and cognitive attributes (e.g., Yan, Ding, Milojević, & Sugimoto, 2012), this study focuses on two other attributes: popularity and impact. It first examines topic dependence though self-citations (Figure 11), then the co-evolving feature of topic popularity and impact between 1964 and 2012 (Figure 12), and finally the correlation relationships among topic popularity, impact, and year of publication (Figure 13).
In Figure 11, one topic has a citation impact of more than 5,000, five between 3,000 and 5,000, 34 between 1,000 and 3,000, and 10 below 1,000. Topics that have the highest citation impact are t1 on user and technology, t4 on scientific collaboration, t16 on citation analysis, t24 on online technologies, and t38 on science literature. Highly cited topics are situated in two broad research areas: informetrics and information systems and technologies.

Similar to disciplines (Yan et al., 2013), research topics also vary to the degree of permeability and can be quantified through topic self-citation ratios. The results show that five topics have self-citation ratios between 0.3 and 0.4, 12 between 0.2 and 0.3, 21 between 0.1 and 0.2, and 12 below 0.1. The most self-contained topics include t8 on GIS and spatial models, t40 on telecommunication, and t4 on scientific collaboration. These topics have a more distinctive core and their papers are published on more specialized subject matters.

In Figure 12, while topic popularity has increased steadily, topic impact reached its peak in 2003 and decreased since then. Because the data set was collected in 2013, the nine-year window indicates that publications in the field of LIS as a whole reached the highest impact nine years after publication. The width of the window is dependent on the number of citable publications of each year and the immediacy of publications being cited.
Because of the nonparametric distribution pattern of citation impact, popularity, and average year of publication, Spearman’s rank correlation coefficient was used to test the relationships of these variables. The result shows that only the correlation between impact and average year of publication is statistically significant $r=0.372$ ($p<0.05$). This indicates that newer topics tend to be cited more often than older topics. This may be attributed to the fact that there is a steady increase of citable publications in recent years. The correlation between impact and popularity is not statistically significant $r=0.279$ ($p=0.056$), which suggests a non-significant overlap between topic popularity and topic impact. Such non-significant relationship also validates the need to have both topic popularity and topic impact in describing topic characteristics. The correlation coefficient between popularity and average year of publication is 0.193 ($p=0.18$), which suggests a non-significant relationship between the age of topics and their likelihood of gaining popularity.
Topical knowledge flow and inheritance

In this subsection, we examine topical knowledge dissemination through the concept of shortest paths. In Figure 14, the most frequently occurred paths are illustrated. Each node represents one topic and is color coded to reflect its average year of publication. The size of a node denotes the number of times a topic occurred on the shortest paths. In total, 511 one-hop knowledge paths are identified. Top 50 paths based on occurrence frequency constitute the primary knowledge path in Figure 14. The frequency ranges from 203 to 29. Eighteen topics are not included by these 50 paths. They are connected through secondary knowledge paths.
Figure 14. Topical knowledge flow. Color coding: average year of publication; size of node: number of times a topic is located on the shortest paths; links: most frequently occurred one-hop paths.

Several knowledge hubs can be identified: t1 on user and technology, t4 on scientific collaboration, t16 on citation analysis, t24 on online technologies, and t38 on science literature play an important role in transferring and dissemination topical knowledge. Three major topical knowledge paths are:

- From electronic resources to digital data and information systems, then to information retrieval systems, then to telecommunication systems, online social networks, health communication, and knowledge management;
- From journal and paper citation impact analysis to collaboration and bibliometrics, then to web science; and
- From collection, literature and reference to health informatics and GIS.

Older topics are largely connected by secondary knowledge paths. This suggests that these topics are no long active in importing or exporting topical knowledge. They are gradually fading away from the
research frontier of LIS. In the meantime, newer topics are tightly connected with the rest of the knowledge flow network. Thus, their topic-level knowledge is expected be incorporated into the knowledge base of LIS in the future.

Conclusion

This study has conducted a topic-level analysis using a data set on library and information science publications. It has highlighted the use of topics in organizing scientific literature and examining research popularity, impact, and dynamics. It has employed a topic modeling technique and conducted a series of data analytics, including dynamic analysis, regression, shortest path, correlation, and similarity analysis. It has found that topics on online technologies, informetrics, information retrieval systems, health communication and informatics, and online social networks have gained popularity over the past decades. Topics related to literature, books, collection, and cataloging, on the other hand, have declined in popularity. These findings, therefore, have addressed the first research question on the dynamic characteristics of topics in LIS. The study has also revealed that LIS journals are becoming more inter-topical. This finding suggests that topic-level studies are necessary to bring more granular perspectives to the existing co-occurrence-based analyses.

By adding citation information, topics with the highest impact have been identified, including those on technology use, collaboration and bibliometrics, citation analysis, and online technologies. This study has also found several topics that are more self-contained than the others, including those on GIS and spatial models, telecommunication, scientific collaboration, and health communication. Through a correlation analysis, this paper has verified a non-significant relationship between topic popularity and topic impact. The result has helped answer the second research question and also validate the need to have both attributes to describe topic characteristics.

Furthermore, a topical knowledge flow network has revealed topic-level knowledge dissemination channels. The study has identified a few major knowledge paths in LIS, including those from electronic resources to online social networks and knowledge management, from journal and paper citation impact to web science, and from collection, literature and reference to health informatics and GIS. It has also found that while older topics are becoming less active in exporting and importing knowledge, newer topics are becoming integral components in the LIS knowledge flow network. These findings have addressed the last research question on the knowledge dissemination patterns of topics in LIS.

This paper has explored topic dynamics and discovered patterns of topic impact and popularity. In order to determine factors that contribute to such dynamic changes, one needs to relate textual data with other data sources (e.g., authorship data and funding information). Future studies will benefit from identifying the mechanism (e.g., shifts in research communities and science policies) that leads to changes in topic impact and popularity.

Acknowledgement

The author would like to thank David McAllister and Yongjun Zhu for their comments to an earlier version of this paper.

References


Leydesdorff, L. (2007). Mapping interdisciplinarity at the interfaces between the Science Citation Index and the Social Science Citation Index. *Scientometrics, 71*(3), 391-405.


