Journal of Informetrics 8 (2014) 98–110

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research dynamics: Measuring the continuity and popularity of research topics

Erjia Yan*

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form 26 October 2013
Accepted 29 October 2013
Available online 23 November 2013

Keywords:
Topic analysis
Networks
Popularity
Continuity
Dynamics

ABSTRACT

Dynamic development is an intrinsic characteristic of research topics. To study this, this paper proposes two sets of topic attributes to examine topic dynamic characteristics: topic continuity and topic popularity. Topic continuity comprises six attributes: steady, concentrating, diluting, sporadic, transforming, and emerging topics; topic popularity comprises three attributes: rising, declining, and fluctuating topics. These attributes are applied to a data set on library and information science publications during the past 11 years (2001–2011). Results show that topics on "web information retrieval", "citation and bibliometrics", "system and technology", and "health science" have the highest average popularity; topics on "*h*-index", "online communities", "data preservation", "social media", and "web analysis" are increasingly becoming popular in library and information science.

1. Introduction

Dynamics is a constant theme in scientific explorations. Research communities may grow or change in size; new species, diseases, or societal patterns may be discovered; and new research topics and specialties may be introduced (Li et al., 2010; Yan, Ding, Milojevic, & Sugimoto, 2012). Over time, some topics are continuously investigated while others appear or disappear (Griffiths & Steyvers, 2004; Upham & Small, 2010; Shi, Nallapati, Leskovec, McFarland, & Jurafsky, 2010). Therefore, it is of great importance to examine research dynamics to understand the evolving cognitive structures of research domains.

Pioneering studies of paper bibliographic coupling networks (Kessler, 1963), paper co-citation networks (Small, 1973), author co-citation networks (White & McCain, 1998), pathfinder networks (White, 2003) and co-word networks (e.g., Callon, Courtial, & Laville, 1991; Ding, Chowdhury, & Foo, 2000; Milojević, Sugimoto, Yan, & Ding, 2011) were capable of identifying research specialties from bibliographic data effectively. However, findings from these studies remained largely static and thus only yielded fixed perspectives on the cognitive structure of research domains.

To examine research dynamics, this study uses a topic modeling technique and proposes two sets of topic attributes–topic continuity and topic popularity. By applying these attributes to a data set of library and information science publications, this study aims to explore the dynamic research landscape of this field for the past 11 years (2001–2011). Specifically, the following questions are addressed:

- How to use topic modeling techniques to study research dynamics?
- What quantitative measurements can be used to describe topic dynamics?
- What topics are present in library and information science? What are their dynamic characteristics?

* Tel.: +1 215 895 1459.
E-mail address: erjia.yan@drexel.edu

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.10.010

This paper provides a complete solution to the analysis of topic dynamics. It allows one to perceive the evolving feature of research topics, the distribution of research specialties, and the dynamic cognitive structure of research fields.

2. Literature review

2.1. Network-based approaches

This subsection reviews the network-based approaches of identifying research topics and specialties. These approaches have been applied to several research levels, including the paper-level (e.g., Chen, 2004, 2006; Kessler, 1963; Small, 1973), the author-level (e.g., Clauset, Newman, & Moore, 2004; White & McCain, 1998; White, 2003), the journal-level (e.g., Glänzel & Schubert, 2003; Leydesdorff & Vaughan, 2006), and the field-level (e.g., Janssens, Zhang, Moor, & Glänzel, 2009; Rafols & Leydesdorff, 2009; Zhang, Liu, Janssens, Liang, & Glänzel, 2010).

Most above-mentioned work used co-occurrence networks as the research instrument. Analyses on lower level research entities, such as papers and authors, usually identified topics and specialties from small but well-defined research fields; whereas analyses on higher level research entities, such as journals and fields, attempted to identify subfields and subdomains from more comprehensive data sets. Both classic clustering techniques (e.g., factor analysis and multidimensional scaling) as well as modern techniques (e.g., edge betweenness, modularity, and hybrid clustering) have been applied.

These studies largely focused on the analysis of research topics and specialties at a single time frame. Thus, results only captured one snapshot of the cognitive structures of chosen fields. Recently, studies have attempted to add dynamic analyses by utilizing multiple time intervals. Several approaches on slicing time intervals are available: intervals that have the same amount of references (e.g., Radicchi et al., 2009), intervals that have the same number of publications (e.g., Sugimoto, Li, Russell, Finlay, & Ding, 2011; Yan & Sugimoto, 2011), same-length intervals (e.g., Åström, 2007; Milojević et al., 2011), and accumulative intervals (e.g., Barabási et al., 2002; Yan & Ding, 2009). These studies laid valuable methodological basis for dynamic analyses of cognitive structures of research fields; however, networks of different time frames were largely analyzed distinctively and a more integrated examination was lacking. In the meantime, empirically, network-based clustering results may require domain expertise to effectively interpret obtained results.

2.2. Topic modeling approaches

Topic modeling techniques use probabilistic models to assign papers, journals, or authors to clusters. A topic can be defined as a probability distribution over terms in a vocabulary (Blei & Lafferty, 2007). Latent Dirichlet Allocation (LDA) model, a classic topic model, was proposed by Blei et al. (2003). The model predicates that words for each paper are derived from a mixture of topics and each topic follows a multinomial distribution.

Since its advent, various modifications of the LDA model have been proposed. As pointed out by Blei and Lafferty (2007), one limitation of the LDA model is that it fails to model topic correlations, which are expected by subsets of a latent topic. Blei and Lafferty (2007) addressed this issue by proposing the correlated topic model (CTM). CTM replaced the Dirichlet prior by a logistic normal distribution which "gives a more realistic model of the latent topic structure" (p. 19). The limitation of CTM is the loss of the conjugacy between the Dirichlet and the multinomial distributions. Thus, simulation approaches, such as Gibbs sampling, may no longer be available.

Tang and colleagues (2008) proposed a unified model called the author-conference-topic (ACT) model to simultaneously model papers, authors, and publication venues. Through ACT model, each author is assigned with a multinomial distribution over topics. Each topic then generates words and determines the assignment of conferences. One recent update of the LDA model is the supervised LDA model. It makes the analyses of multi-labeled corpora (e.g., tags from delicious.com and various classifications) possible. Blei and McAuliffe's (2010) version of supervised LDA can successfully address this challenge, but a document can only be assigned with one label. Ramage, Hall, Nallapati, and Manning (2009) offered an approach which enabled the multi-label assignment. Their supervised labeled LDA (L-LDA) associated one label with one topic and allowed the model to learn word-label relations.

Through topic modeling techniques, topic dynamics has been examined mainly through the following approaches: post hoc analysis (e.g., Griffiths & Steyvers, 2004; Hall, Jurafsky, & Manning, 2008), segmented approaches (e.g., Bolelli, Ertekin, Zhou, & Giles, 2009), and continuous-time model (Wang & McCallum, 2006). Post hoc analysis uses topic-document probability distributions to evaluate the presence of identified topics. Segmented approaches build the dynamic component in the probabilistic model. It assumes that the state of topics at a single time point is independent from all other time points and divides document corpora into segments that have contingent time stamps (Bolelli et al., 2009). Continuous-time model is a non-Markov model proposed by Wang and McCallum (2006), where they found the non-Markov model provides better prediction and more interpretable topical trends. In this study, a post hoc dynamic analysis using the ACT model is selected because of its marked performance (Tang et al., 2008) as well as its advanced input and output support.

3. Methods

3.1. Identifying topics

Topic dynamics is calculated through the Author-Conference-Topic (ACT) model (Tang et al., 2008). Specifically, θ_i is the topic distribution for document i. Mean θ ($\bar{\theta}$), therefore, is a direct quantitative measurement to assess topic popularity: the higher the $\bar{\theta}$, the more visible the topic, and thus the more popular that topic is (Griffiths & Steyvers, 2004). For instance, assuming the number of topics is set at ten, $\bar{\theta}$ of 0.1 means this topic has an average popularity. A value above 0.1 suggests a more visible and thus more popular topic and a value below 0.1 suggests a less visible topic. Because the data set spans 11 years, 11 independent ACT models were run, one for each year of the data set based on year of publication.

For topics from the same time interval, the calculation of topic similarity can be calculated directly as they share the same array of words. However, for topics from different time intervals, extra steps are needed to calculate topic similarity for post hoc dynamic analysis. The procedures are: first, the union of all unique words for all time intervals was obtained; then, for those words that did not show up in certain time intervals, their word-topic distribution (P(word | topic)) was set with zeros. Therefore, topics from different time periods contained the same array of words.

The Jensen–Shannon divergence (JSD) was used as the similarity measurement to quantify the topic similarity between different word-topic distributions. JSD has been widely used in the text mining community to evaluate the information entropy among various probabilistic distributions. It is a more effective metric than geometrical-based measurements (Lee, 2001). JSD is a symmetrized and smoothed version of the Kullback–Leibler divergence (KLD). For instance, JSD between two subjects P and Q can be presented as: $JSD(P||Q) = 1/2D(P||M) + 1/2D(Q||M)$, where $M = 1/2D(P+Q)$, and $D(P||M)$ is the KLD between two probability distributions P and M: $D(P||M) = \sum_i P(i)\ln p(i)/M(i)$ for discrete random variables and $D(P||M) = \int P(x)\ln p(x)/M(x)dx$ for continuous random variables. The upper bound of JSD is In(2) in this study. As a divergence measure, the smaller the JSD, the higher the similarity is.

3.2. Dynamic topic characteristics

Using the topic popularity measurement $\bar{\theta}$ and Jensen–Shannon divergence, two sets of topic attributes can thus be defined: topic continuity and topic popularity. Topic continuity comprises six attributes: steady, concentrating, diluting, sporadic, transforming, and emerging topics. Topic popularity comprises three attributes: rising, declining, and fluctuating. Topic continuity and popularity attributes are formally defined in the following paragraphs.

JSD between topics from two adjacent time intervals is calculated first, thus forming a JSD matrix. The procedures are:where n is the number of topics and i indicates topics from the former time interval and j indicates topics from the

```
for i = 1:n
        for j = 1:n
                P(word | topic_i) = P(word | topic_i)/SUM(P(word | topic_i))
                P(word | topic_j) = P(word | topic_j)/SUM(P(word | topic_j))
                d = JSD(P(word | topic_i), P(word | topic_j));
                JSD_Matrix(i,j) = d;
        end
end
```

latter time interval. Fig. 1 shows a heatmap representation of JSD matrices comprising ten topics.

In order to track the same topic from two adjacent time intervals, the minimum value for each row of a JSD matrix was used, referred to as the joint JSD score (JJSDS): MIN(JSD_Matrix(i,j)), for $j = 1{:}n$. For instance, in the left panel of Fig. 1, the first topic of time slice 1 is developed into the ninth topic of time slice 2. Applying the same approach to each pair of adjacent time slices, for each topic, an array of JJSDS can be obtained. Because for a given topic, JJSDS identifies only one topic in the following time interval, multiple topic assignment (i.e., topic divergence) thus could not be included.

Situating an array of JJSDS in the context of a single independent variable linear regression, the slope can be obtained. For purpose of an effective comparison, slopes were then normalized through the z score: $z(slope_{\text{topic}_i}) = (slope_{\text{topic}_i} - \sum_{i=1}^{n} slope_{\text{topic}_i})/n/SD(slope_{\text{topic}_{1:n}})$ where n is the number of topics that spanned more than five years. z scores were used to quantify topic continuity attributes.

Steady: $-0.5 \le z(slope_{\text{topic}_i}) \le 0.5$.

Concentrating: $z(slope_{\text{topic}_i}) < -0.5$.

Diluting: $z(slope_{\text{topic}_i}) > 0.5$.

z Scores of −0.5 and 0.5 were chosen because these two scores are associated with probabilities in normal distribution that may be convenient to categorize topics: −0.5 or below corresponds to 30% which is the percentage of diluting topics; 0.5 or above corresponds to another 30% which is the percentage of concentrating topics; the remaining 40% is the percentage of steady topics. For real-world distributions, these probabilities may be different; yet, these z scores can be used consistently

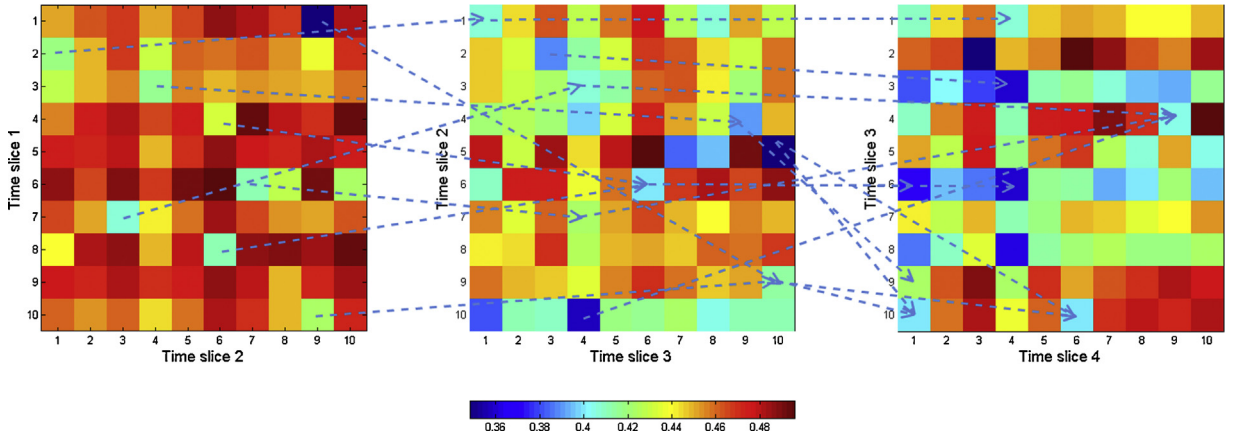

Fig. 1. Heatmap based on Jensen–Shannon divergence scores (lines suggest topic connections).

as guidelines to label topics. Different z scores may also be applied to threshold the percentage of topics that fall into each category; its discussion, however, is not in the scope of the current study.

Sporadic: $z(JJSDS_{\text{topic}_{i:t-1\to t}}) > 0 \&\& z(JJSDS_{\text{topic}_{i:t:t+1}}) > 0$, for topic i, where $z(JJSDS_{\text{topic}_{i:t\to t+1}}) = (JJSDS_{\text{topic}_{i:t\to t+1}} - \sum_{a=1}^{n} JSD_a)/n/SD(JSD_a)$, and n is the total number of JSD (in this case, $n = 4000$: 20 topics × 20 topics × 10 intervals). If the z score of a JJSDS is larger than 0, it suggests that this topic connection is weaker than the average JSD and thus the connected topics may not necessarily carry common topical characteristics.

Transforming: $\exists j : -2 < z(JJSDS_{\text{topic}_{j:t\to i:t+1}}) \le 0$, for topic i. It means that a related but not identical topic j (as characterized by the z score) is added to topic i. Topic i is thus *transformed* by containing both topical characteristics from i and j.

Emerging: $t : JJSDS_{\text{topic}_{i:t-1\to t}} \&\& j : JJSDS_{\text{topic}_{j:t\to i:t+1,}}$, for $t > 2$, topic i. It means topic i emerged in year t and no predecessors can be found, and at the same time, no other topic such as topic j is transformed into topic i at $t+1$. The combination of these two rules guarantees that the emerging topic is not a close variation of other topics.

The attributes of steady, concentrating, and diluting topics focus on the overall topical characteristics whereas the attributes of sporadic, transforming, and emerging topics focus on the topical characteristics of a specified time frame. Therefore, these attributes are not mutual exclusive, suggesting that a topic can be a concentrating topic overall, and in the meantime, related topics were added and thus qualifying it for a transforming topic. Fig. 2 conceptualizes the six topic continuity attributes.

For each array of JJSDS, its corresponding mean θ scores ($\bar{\theta}$) across different time slices can be extracted. Because these scores are from different distributions, they may have varied distributions patterns. To overcome this, these mean θ scores were normalized though the z score in relation to their individual yearly distributions: $z(\bar{\theta}_{\text{topic}_i,t})(\bar{\theta}_{\text{topic}_i,t} - \sum_{i=1}^{n} \bar{\theta}_{\text{topic}_i,t}/n)/SD(\bar{\theta}_{\text{topic}_{i:n},t})$, where n is the number of topics and t is a particular year. These arrays of z scores are used

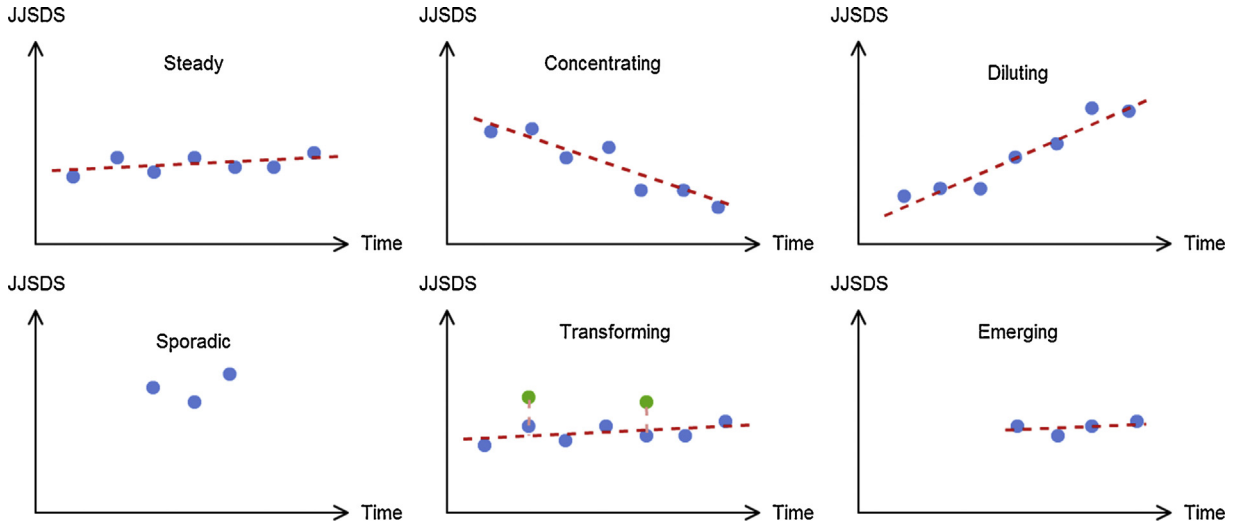

Fig. 2. An illustration of topic continuity attributes.

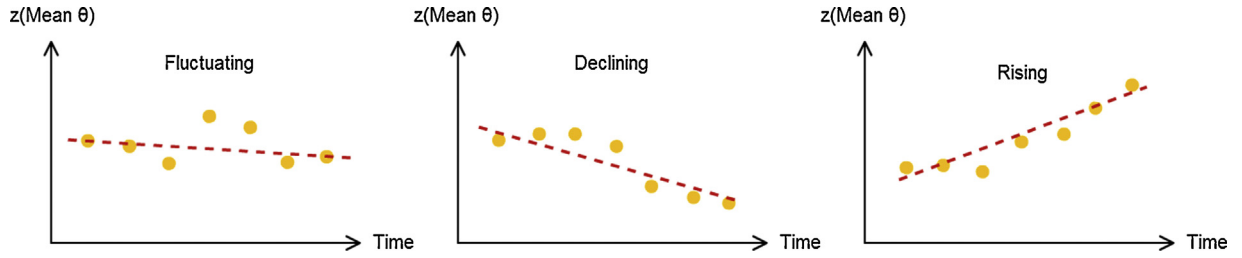

Fig. 3. An illustration of topic popularity attributes.

Table 1
Data statistics.

Publication year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
No. of publications	1985	2039	2076	2058	2427	2338	2526	2440	2411	3830	3666

to describe topic popularity attributes, also in the context of a single independent variable linear regression. Slopes from these linear regressions were obtained and normalized through $z(slope_{\text{topic}_i})$, using the same method as mentioned in the preceding paragraph. The dependent variable for popularity attributes is $z(\bar{\theta}_{\text{topic}_i,t})$, which is different from the continuity attributes in that the dependent variable is JJSDS, because as a similarity measure, JJSDS itself is standardized and no more normalization is necessary for input of a linear regression.

Rising: $z(slope_{\text{topic}_i}) > 0.5$.

Declining: $z(slope_{\text{topic}_i}) < -0.5$.

Fluctuating: $-0.5 \leq z(slope_{\text{topic}_i}) \leq 0.5$.

The −0.5 and 0.5 thresholds were chosen for the same reason that each category will have comparable amount of topics. Fig. 3 conceptualizes the three topic popularity attributes.

3.3. Data set

The data set contains publications of all journals indexed in the 2011 version of the Journal Citation Report in the Information Science & Library Science subject category. Articles, proceeding papers, and review articles published within these journals from 2001 to 2011 were downloaded for analysis (downloading time: October 2012). Stop words were then removed from publications' titles.[1] Publications without titles, authors, or journal names were removed from the data set. The final data set comprised 27,796 papers. Table 1 shows the yearly distribution of papers.

The number of topics is set at 20: this number considers the size of the paper corpus as well as previous empirical studies on the cognitive structure of library and information science (e.g., Milojević et al., 2011; Sugimoto et al., 2011; White & McCain, 1998; Zhao & Strotmann, 2008). For reasons of consistency, the same number of topics was identified for each year of the data set. It is a standard approach for longitudinal studies of research topics (e.g., Griffiths & Steyvers, 2004; Hall et al., 2008; Shi et al., 2010; Sugimoto et al., 2011).

4. Results

4.1. Topics in library and information science

In this subsection, we first present histograms made from values in Jensen–Shannon divergence (JSD) matrices (Fig. 4). These histograms provide a direct perception on how research topics in library and information science are related as measured by JSD. This subsection then introduces all 20 topics in each year from 2001 to 2011 as well as how topic continuity and popularity attributes are applied to these topics (Fig. 5).

Fig. 4 uses histograms to visualize JSD values for each pair of adjacent years. Because there are 20 topics for each year, the number of data points in each histogram is 400 (20×20). This number is 4000 for the histogram in the lower right section of Fig. 4, as it uses JSD values for all pairs of adjacent years.

Although histograms in Fig. 4 are seemingly normal distributed, all histograms rejected the null hypothesis that the data points come from a standard normal distribution in one-sample Kolmogorov–Smirnov test ($p < 0.05$). The result can mainly be attributed to the outliers on the left side of the x-axis in each histogram. Most of these outliers are precisely the joint JSD scores (JJSDS), which connects topics in adjacent years. While the mean for all JSD values is 0.42, the mean of individual histograms have reduced gradually from 0.44 in 2001/2002 to 0.41 in 2010/2011. Such results signify more

[1] The stop word list used is at: http://www.pages.drexel.edu/~ey86/p/lis_topic/stoplist.txt.

[3] High definition images can be downloaded at: http://www.pages.drexel.edu/~ey86/p/lis_topic/.

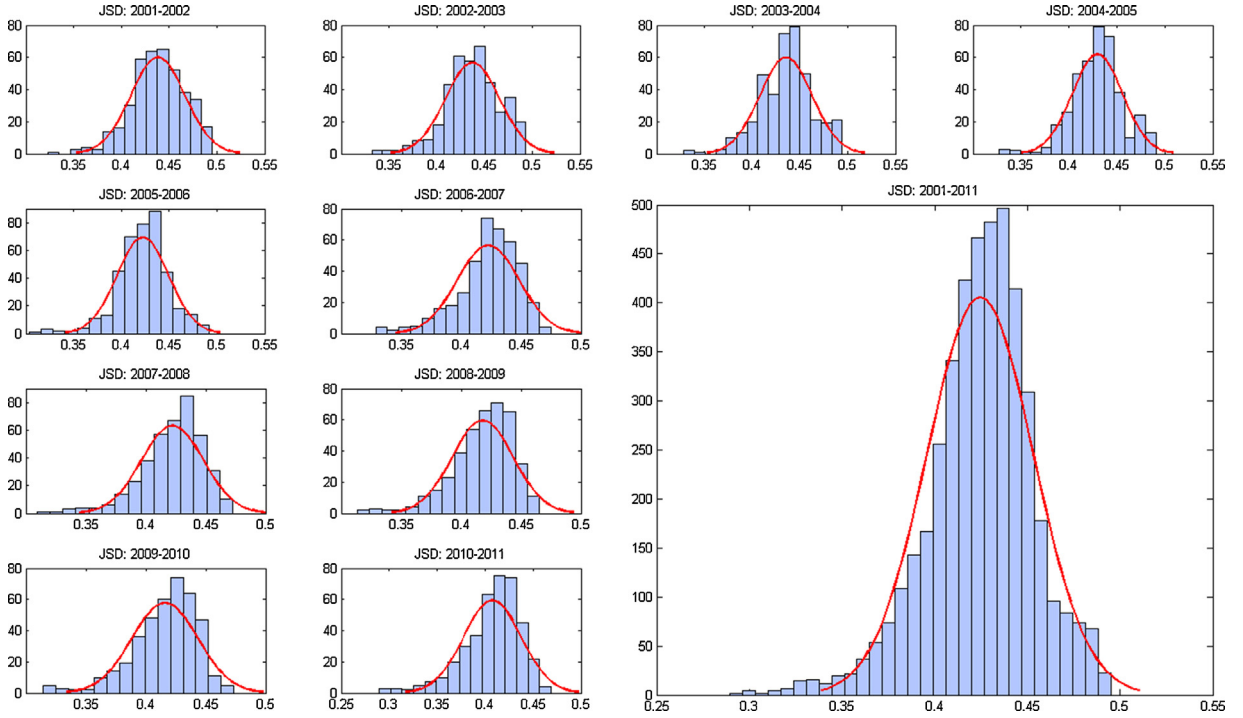

Fig. 4. Histograms of JSD values (from 2001 to 2011).

coherent foci of research conducted in recent years in library and information science and may also imply a more inter-topical or inter-specialty research trend.

Fig. 5 is the principal visualization of this study: it lists top five words that are most likely to be associated with each topic. It highlights different types of connection strength through the z score (z(JJSDS)) as well as different types of topics. In total, 66 topics can be identified (as marked on the figure) and among them 49 topics span five or more years. Topic continuity and popularity attributes are then applied to these 49 topics (because slope will no longer be reliable for data points less than four).

Because a topic at time t finds one succeeding topic at time $t+1$, 20 connections (i.e., JJSDS) are available between two topics of adjacent years. Among all 180 connections, 77 (43%) have a z(JJSDS) <-2; 68 (38%) have a $-2 < z$(JJSDS) $<= -1$;

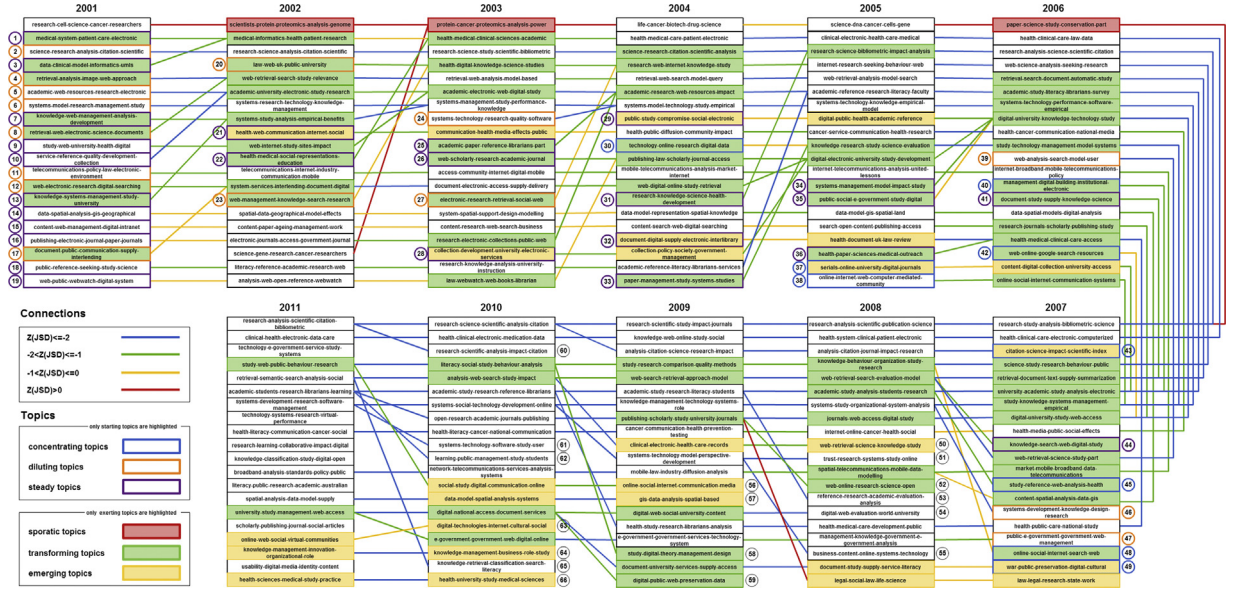

Fig. 5. Topics in library and information science (an overview).[3]

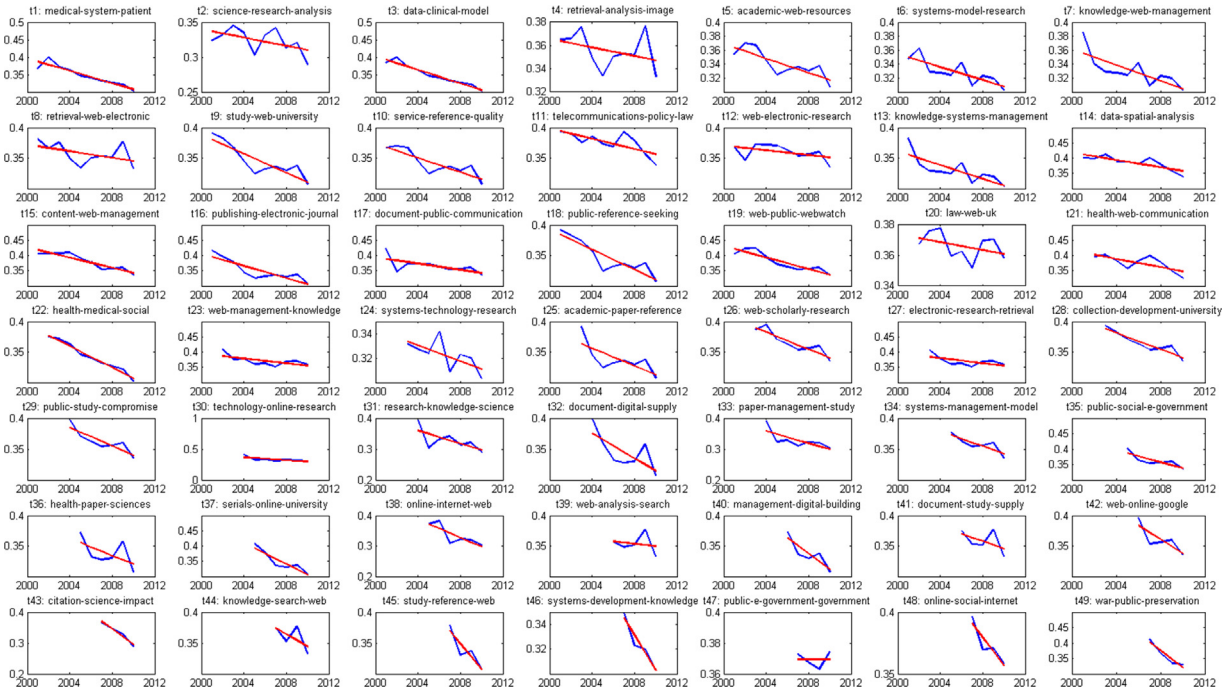

Fig. 6. Dynamics of topic continuity for 49 topics.

28 (15%) have a $-1 < z(\text{JJSDS}) <= 0$; and 7 (4%) have a $z(\text{JJSDS}) > 0$. The result shows that most JJSDS have a below-average divergence (JSD) and thus possessing an above-average similarity.

Among all the 49 topics that last for five or more years, 25 topics are steady topics, 15 are diluting topics, and 9 are concentrating topics. Additionally, there are 29 instances of transforming topics and 13 instances of emerging topics. These topic continuity attributes will be elaborated in the following paragraphs.

4.2. Topic continuity and popularity

In this subsection, we first present the results on dynamics of topic continuity (Fig. 6) and topic popularity (Fig. 7); we then introduce a list of all topics with their *z* scores of slopes on topic continuity and popularity (Fig. 8); transforming topics (Fig. 9) and emerging topics (Fig. 10) are also detailed in this subsection.

Fig. 6 shows the line chart of topic continuity for each topic (*y*-axis: JJSDS). Due to the space limitation, only the top three words that have the strongest topical association are displayed (instead of five words in Fig. 5). Data fitting curves are canvassed in red.

Based on an observation of the data fitting curves, in an absolute sense, all topics are becoming somewhat more concentrated (slope < 0). These slopes were then normalized through the *z* score and thus, in a relative sense, steady, concentrating, and diluting topics can then be identified. The relative comparison is regarded as more useful in perceiving topical characteristics, as similar to many real-life activities that co-evolve over time, topics continuity and popularity also have the propensity to change. Thus, it is not about *whether* they change but *how* they change; changing patterns can be more effectively examined through a relative lens.

Several topics possess a smoother continuity trend, such as t1: medical-system-patient, t3: data-clinical-model, t14: data-spatial-analysis, t15: content-web-management, t22: health-medical-social, t30: technology-online-research, t43: citation-science-impact, t46: systems-development-knowledge, and t49: war-public-preservation. Meanwhile, the continuity trend for a few topics is more rugged, including t2: science-research-analysis, t4: retrieval-analysis-image, and t20: law-web-uk.

Fig. 7 shows the line chart of topic popularity for each topic (*y*-axis: $z(\bar{\theta})$). The calculation of popularity is limited to the data set. As research is becoming more interdisciplinary, an increasing amount of scholarly journals are indexed in multiple subjects; therefore, a topic may have a low presence in the current data set – resulting in a low popularity, but it may be a popular topic in other subjects/data sets.

Compared with slopes of topic continuity, slopes of topic popularity are more diversified: some are smaller than zero (e.g., t1, t3, and t34), some very close to zero (e.g., t5, t6, t7, t8, t9, and t10), and some larger than zero (e.g., t2, t14, t33, t39, t43, t47, t48, and t49). For purpose of an effective comparison, these slopes were also normalized through the *z* score. While most topics have a rugged popularity trend, a few are smoother than the other, for instance, t8: retrieval-web-electronic, t31: research-knowledge-science, t43: citation-science-impact, and t48: online-social-internet.

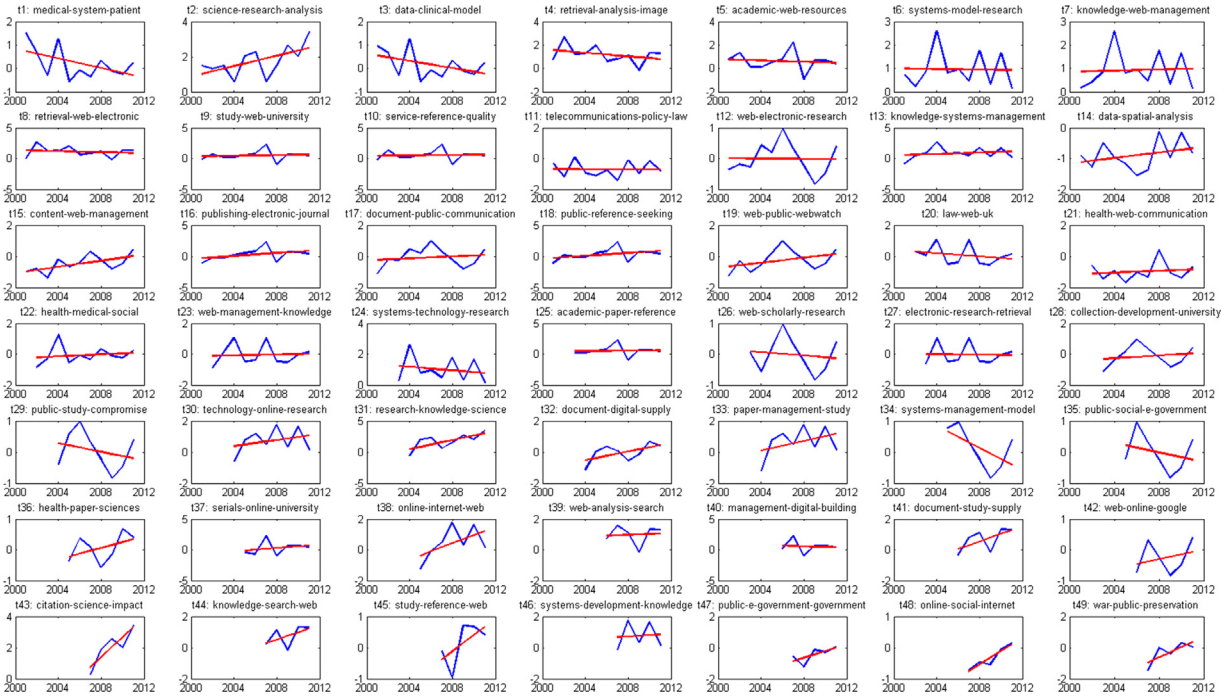

Fig. 7. Dynamics of topic popularity for 49 topics.

The normalized slopes of continuity and popularity are listed in Fig. 8. This figure also includes the starting year for each topic and the mean and rank measured by mean JJSDS and $z(\bar{\theta})$ respectively.

Topics with the most significant concentrating trend are t49: war-public-preservation-digital-web, t43: citation-science-impact-scientific-index, t45: study-reference-web-analysis-health, t37: serials-online-university-digital-journal, and t38: online-internet-web-mediated-community, indicating that research on these topics is increasingly becoming specialized. Topics with the most significant diluting trend are t47: public-e-government-government-web-management, t20: law-web-uk-public-university, t4: data-spatial-analysis-gis-geographical, t39: web-analysis-search-model-user, and t12: web-electronic-research-digital-searching, suggesting that these topics are becoming more permeable and inclusive.

In regards to topic popularity, topics with the most significant rising trend are t43: citation-science-impact-scientific-index, t31: research-knowledge-science-health-development, t49: war-public-preservation-digital-cultural, t38: online-internet-web-computer-mediated-community, and t45: study-reference-web-analysis-health. These topics are characterized by the concepts of "citation", "online", and "health", indicating that these research concepts are becoming prevalent in library and information science. Topics with the most significant declining trend are t34: systems-management-model-impact-study, t1: medical-system-patient-care-electronic, t3: data-clinical-model-informatics-umls, t35: public-social-e-government-study-digital, and t4: retrieval-analysis-image-web-approach. These topics are characterized by the concepts of "systems" and "model", suggesting that they are becoming less explored in library and information science (but may still be the research foci outside library and information science).

The rising topics of t43: citation-science-impact-scientific-index, t2: science-research-analysis-citation-scientific, and t31: research-knowledge-science-health-development also possess the highest average popularity ($z(\bar{\theta})$), indicating that these topics were visible to begin with and are becoming even more popular in recent years. Based on their $z(\bar{\theta})$, two retrieval related topics t4: retrieval-analysis-image-web-approach and t8: retrieval-web-electronic-science-documents were highly visible; however, the trend suggests that they are becoming less prevalent in library and information science in recent years. Such a trend may be attributed to the shift of scholarly communication channel of information retrieval research from information science venues to computer science venues.

Relating continuity trend and popularity trend, it is found that topics with rising popularity tend to be concentrating ones and topics with declining popularity tend to be diluting ones. This relationship is statistically significant with Spearman's rank correlation coefficient $r = -0.48$, $p < 0.01$. The association between mean JJSDS and $z(\bar{\theta})$ is even stronger, with $r = -0.79$, $p < 0.01$.

In addition to topic continuity attributes of steady, concentrating, and diluting, transforming topics (Fig. 9) and emerging topics (Fig. 10) are also introduced here.

Transforming topics are characterized by the consolidation of a few related but not identical topics in the preceding year to one topic in the following year. For instance, two health related topics in 2001, medical-system-patient-care-electronic and data-clinical-model-informatics-umls, have been transformed into medical-informatics-health-patient-research in 2002

			Continuity		Popularity	
ID	Year	Topics	Dynamics	M (Rank)	Dynamics	M (Rank)
1	2001	medical-system-patient-care-electronic	steady (Z(slope)=-0.09)	0.347 (29)	declining (Z(slope)=-1.22)	0.211 (25)
2	2001	science-research-analysis-citation-scientific	diluting (Z(slope)=0.96)	0.324 (48)	raising (Z(slope)=0.52)	1.752 (2)
3	2001	data-clinical-model-informatics-umls	steady (Z(slope)=-0.24)	0.349 (27)	declining (Z(slope)=-1.04)	0.160 (26)
4	2001	retrieval-analysis-image-web-approach	diluting (Z(slope)=1.16)	0.355 (23)	declining (Z(slope)=-1.04)	1.172 (4)
5	2001	academic-web-resources-research-electronic	diluting (Z(slope)=0.57)	0.340 (34)	declining (Z(slope)=-0.70)	0.613 (16)
6	2001	systems-model-research-management-study	diluting (Z(slope)=0.66)	0.329 (44)	declining (Z(slope)=-0.55)	0.958 (8)
7	2001	knowledge-web-management-analysis-development	steady (Z(slope)=0.47)	0.330 (42)	fluctuating (Z(slope)=-0.43)	0.926 (9)
8	2001	retrieval-web-electronic-science-documents	diluting (Z(slope)=1.01)	0.357 (22)	declining (Z(slope)=-0.79)	1.100 (5)
9	2001	study-web-university-health-digital	steady (Z(slope)=0.11)	0.345 (30)	fluctuating (Z(slope)=-0.31)	0.491 (20)
10	2001	service-reference-quality-development-collection	steady (Z(slope)=0.44)	0.342 (33)	fluctuating (Z(slope)=-0.39)	0.525 (19)
11	2001	telecommunications-policy-law-electronic-environment	diluting (Z(slope)=0.74)	0.375 (4)	declining (Z(slope)=-0.52)	-0.707 (47)
12	2001	web-electronic-research-digital-searching	diluting (Z(slope)=1.13)	0.360 (18)	declining (Z(slope)=-0.53)	-0.008 (32)
13	2001	knowledge-systems-management-study-university	steady (Z(slope)=0.49)	0.330 (43)	fluctuating (Z(slope)=-0.09)	0.831 (10)
14	2001	data-spatial-analysis-gis-geographical	steady (Z(slope)=0.42)	0.384 (1)	fluctuating (Z(slope)=-0.20)	-0.896 (48)
15	2001	content-web-management-digital-intranet	steady (Z(slope)=0.00)	0.380 (2)	fluctuating (Z(slope)=0.18)	-0.484 (46)
16	2001	publishing-electronic-journal-paper-journals	steady (Z(slope)=-0.29)	0.351 (25)	fluctuating (Z(slope)=0.32)	0.277 (23)
17	2001	document-public-communication-supply-interlending	diluting (Z(slope)=0.58)	0.365 (11)	fluctuating (Z(slope)=-0.29)	-0.077 (39)
18	2001	public-reference-seeking-study-science	steady (Z(slope)=0.02)	0.348 (28)	fluctuating (Z(slope)=0.26)	0.293 (22)
19	2001	web-public-webwatch-digital-system	steady (Z(slope)=-0.19)	0.379 (3)	fluctuating (Z(slope)=0.06)	-0.259 (41)
20	2002	law-web-uk-public-university	diluting (Z(slope)=1.26)	0.366 (10)	declining (Z(slope)=-0.87)	0.064 (29)
21	2002	health-web-communication-internet-social	steady (Z(slope)=0.29	0.375 (5)	fluctuating (Z(slope)=-0.30)	-0.979 (49)
22	2002	health-medical-social-representations-education	steady (Z(slope)=-0.07)	0.342 (32)	fluctuating (Z(slope)=-0.28)	-0.069 (38)
23	2002	web-management-knowledge-search-research	diluting (Z(slope)=0.77)	0.371 (7)	fluctuating (Z(slope)=-0.42)	-0.052 (37)
24	2003	systems-technology-research-quality-software	diluting (Z(slope)=0.91)	0.323 (49)	declining (Z(slope)=-0.92)	1.005 (6)
25	2003	academic-paper-reference-librarians-part	steady (Z(slope)=0.20)	0.338 (36)	fluctuating (Z(slope)=-0.39)	0.525 (18)
26	2003	web-scholarly-research-academic-journal	steady (Z(slope)=0.25)	0.365 (12)	declining (Z(slope)=-0.69)	-0.024 (34)
27	2003	electronic-research-retrieval-social-web	diluting (Z(slope)=0.75)	0.370 (8)	declining (Z(slope)=-0.55)	-0.038 (35)
28	2003	collection-development-university-electronic-services	steady (Z(slope)=0.28)	0.364 (13)	fluctuating (Z(slope)=-0.20)	-0.138 (40)
29	2004	public-study-compromise-social-electronic	steady (Z(slope)=0.14)	0.362 (14)	declining (Z(slope)=-0.99)	0.039 (31)
30	2004	technology-online-research-digital-data	concentrating (Z(slope)=-0.53)	0.330 (41)	fluctuating (Z(slope)=0.18)	0.717 (13)
31	2004	research-knowledge-science-health-development	steady (Z(slope)=-0.39)	0.328 (46)	raising (Z(slope)=2.00)	1.726 (3)
32	2004	document-digital-supply-electronic-interlibrary	steady (Z(slope)=-0.30)	0.345 (31)	fluctuating (Z(slope)=0.46)	-0.040 (36)
33	2004	paper-management-study-systems-studies	steady (Z(slope)=-0.26)	0.328 (45)	raising (Z(slope)=0.56)	0.636 (15)
34	2005	systems-management-model-impact-study	steady (Z(slope)=0.41)	0.358 (21)	declining (Z(slope)=-1.78)	0.124 (28)
35	2005	public-social-e-government-study-digital	steady (Z(slope)=-0.24)	0.362 (15)	declining (Z(slope)=-1.04)	-0.017 (33)
36	2005	health-paper-sciences-medical-outreach	steady (Z(slope)=0.23)	0.338 (37)	fluctuating (Z(slope)=0.15)	0.060 (30)
37	2005	serials-online-university-digital-journals	concentrating (Z(slope)=-1.68)	0.349 (26)	fluctuating (Z(slope)=0.44)	0.273 (24)
38	2005	online-internet-web-computer-mediated-community	concentrating (Z(slope)=-1.14)	0.335 (39)	raising (Z(slope)=1.36)	0.437 (21)
39	2006	web-analysis-search-model-user	diluting (Z(slope)=1.14)	0.353 (24)	fluctuating (Z(slope)=-0.34)	0.981 (7)
40	2006	management-digital-building-institutional-electronic	concentrating (Z(slope)=-0.82)	0.337 (38)	declining (Z(slope)=-0.89)	0.540 (17)
41	2006	document-study-supply-knowledge-science	steady (Z(slope)=0.44)	0.358 (20)	raising (Z(slope)=1.21)	0.670 (14)
42	2006	web-online-google-search-resources	concentrating (Z(slope)=-0.56)	0.360 (17)	fluctuating (Z(slope)=0.03)	-0.264 (42)
43	2007	citation-science-impact-scientific-index	concentrating (Z(slope)=-2.98)	0.333 (40)	raising (Z(slope)=3.99)	2.027 (1)
44	2007	knowledge-search-web-digital-study	steady (Z(slope)=-0.29)	0.359 (19)	raising (Z(slope)=1.12)	0.765 (11)
45	2007	study-reference-web-analysis-health	concentrating (Z(slope)=-2.17)	0.338 (35)	raising (Z(slope)=1.33)	0.144 (27)
46	2007	systems-development-knowledge-design-research	diluting (Z(slope)=-1.05)	0.324 (47)	fluctuating (Z(slope)=-0.21)	0.765 (12)
47	2007	public-e-government-government-web-management	diluting (Z(slope)=1.49)	0.369 (9)	raising (Z(slope)=1.01)	-0.439 (45)
48	2007	online-social-internet-search-web	concentrating (Z(slope)=-0.52)	0.374 (6)	raising (Z(slope)=0.99)	-0.328 (44)
49	2007	war-public-preservation-digital-cultural	concentrating (Z(slope)=-3.48)	0.361 (16)	raising (Z(slope)=1.79)	-0.302 (43)

Continuity	concentrating topics Z(slope:JJSDS)<-0.5	steady topics -0.5=<Z(slope:JJSDS)<=0.5	diluting topics Z(slope:JJSDS)>0.5
Popularity	declining topics Z(slope:θ)<-0.5	fluctuating topics -0.5=<Z(slope:θ)<=0.5	raising topics Z(slope:θ)>0.5

Fig. 8. List of topics and their continuity and popularity attributes.

which comprises both core concepts of "patient" and "informatics"; three research collection related topics in 2003, academic-paper-reference-librarians-part, academic-electronic-web-digital-study, and research-electronic-collections-public-web, haven been transformed into academic-research-web-resources-impact in 2004 which comprises core concepts of "academic", "research", and "web". In the meantime, a few topics have been continuously developed; for instance, two retrieval related topics in 2006, retrieval-search-document-automatics-study and document-study-supply-knowledge-science, have been developed into retrieval-documents-text-supply-summarization in 2007; this very topic was further developed into web-retrieval-search-evaluation-model in 2008 by incorporating two other "web" related topics in 2007: web-retrieval-science-study-part and knowledge-search-web-digital-study.

Thirteen emerging topics were identified through applying topic continuity attributes to the data set (Fig. 10).

Fig. 9. Transforming topics in library and information science.

Key terms can be detected from these top words that have the strongest topical associations, such as “health”, “data”, “web”, “social”, “online”, and “impact”. The first seven topics in Fig. 10 also have a popularity measure in Fig. 8. By relating these two figures, we find that except for t43: citation-science-impact-scientific-index, other topics have an average popularity ($z(\bar{\theta})$) smaller than the average, suggesting that they are not deemed as popular topics yet; trend-wise, most of them have a popularity attribute of fluctuating, which calls for a future examination of the status of these topics. Topic 43, the one

		Emerging year
21	health-web-communication-internet-social	2002
28	collection-development-university-electronic-services	2003
29	public-study-compromise-social-electronic	2004
32	document-digital-supply-electronic-interlibrary	2004
37	serials-online-university-digital-journals	2005
43	citation-science-impact-scientific-index	2007
49	war-public-preservation-digital-cultural	2007
50	web-retrieval-science-knowledge-study	2008
56	online-social-internet-communication-media	2009
57	gis-data-analysis-spatial-based	2009
63	digital-technologies-internet-cultural-social	2010
64	knowledge-management-business-role-study	2010
66	health-university-study-medical-sciences	2010

Fig. 10. Emerging topics in library and information science.

on *h*-index[2] and citation analysis, is the most popular topic. This topic first emerged in 2007, which is two years after the first *h*-index paper by Hirsch. This delay is expected considering the publication cycle. Some of these emerging topics were developed into other topics. For instance, t43 on *h*-index was developed as research-analysis-scientific-citation-bibliometric in 2011; t50 on web retrieval was developed as retrieval-semantic-search-analysis-social in 2011; t56 on social media was developed as study-web-public-behavior-research in 2011.

5. Discussion

5.1. Comparison with related studies

Research topics have the propensity to evolve. Alteration of the ever shifting patterns is redolent of evolutionary challenge. This is especially evident in fast growing fields. This study has identified the dynamic characteristics of research topics in library and information science from the perspective of topic continuity and topic popularity.

This study finds that in library and information science, research topics on "web information retrieval", "citation and bibliometrics", "system and technology", and "health science" have the highest average popularity over the past decade (from 2001 to 2011). Research on "*h*-index", "online communities", "data preservation", "social media", and "web analysis" are increasingly becoming popular topics.

Overall, findings of this study are consistent with previous studies using co-word, co-citation, and topic modeling techniques. For instance, a co-word study by Milojević and colleagues (2011) has found that title terms "citation", "impact factor", and "web" have a rising usage from 1989 to 2008. Other related dynamic studies that cover the target time frame of the current study (2001–2011) include Åström's (2007) study on examining library and information science research front, where the study found that webometrics and information-seeking and retrieval have become dominating research areas between 2000 and 2004. This finding has been verified by Klavans and Boyack (2011) where the authors used the global map (i.e., the map of science) to enhance to accuracy of local maps (i.e., the contextual map of information science). They identified five core areas in information science, including information-seeking behavior, computer-enhanced retrieval, scientometrics, co-citation analysis, and citation behavior. Besides the contextual analysis of information science, structural analysis has also been achieved from a time-series empowered author co-citation and document co-citation analysis (Chen, Ibekwe-SanJuan, & Hou, 2010). Through the application of a series of structural metrics such as centrality measures, modularity and silhouette, a clear cognitive structure of information science was attained in that the research areas of interactive information retrieval, academic web, information retrieval, citation behavior, and *h*-index have gained a particular popularity from 1996 to 2008. In addition to journal publications, Sugimoto and colleagues (2011) applied a LDA model to library and information science dissertations and demonstrated dissertations as an important communicative genre. Their study indicated that between 2000 and 2009, internet and information retrieval related topics were the central dissertation research themes. Above mentioned have identified that information retrieval, web studies, and bibliometrics and citation analysis are the predominant research areas in information science; their dynamics, however, were not coherently explored. The contribution of the current study is that it proposes two sets of quantitative topic attributes. These attributes have streamlined the dynamic analysis of research topics and specialties and have further complemented co-occurrence-based studies.

5.2. Limitations

This paper has identified dynamic characteristics of topics in library and information science; however, limited information can be told about the mechanisms that resulted in such characteristics. That being said, the study is unable to pinpoint, for instance, whether the growing popularity of network and citation studies is the result of a growing research community, a drive by the commercial market, a stimulus from funding agencies, or a combination of these or other unlisted factors. Popular topics may be associated with research communities that are expanding in size and/or tend to have higher productivity. Conversely, less popular topics may be associated with communities that are shrinking and/or have a reduced productivity. Topic continuity and popularity attributes reflect research specialties' development in scientific communities, which is further guided by science policies and the attention of the general public. Future analysis on this direction may benefit from revealing the causal mechanisms that potentially shape topic characteristics (e.g., Ding, Yan, Sugimoto, & Milojevic, 2013; Shi et al., 2010).

6. Conclusion

In informetrics, studies have mainly focused on analyzing the performance and the social and cognitive implications of several types of research entities, including papers, authors, institutions, journals, and fields. Authors and institutions are typically used to examine social relations in academia; while journals and fields are predominantly used to investigate the cognitive structure of research domains. Although journals and fields provide a fixed view of the cognitive

[2] *h*-index for some papers was indexed as h index in Web of Science and because the standalone letter "*h*" is a stopword, only "index" is left. For instance, two articles by Hirsch used "index *h*" and "*h* index" without hyphen.

structure of a variety of domains, a more granular perspective may be prudent. Topic analysis can precisely provide a more refined assessment by clustering research papers based on certain probability distributions. Because of such quantitative results, a more integrated dynamic cognitive analysis is thus possible, as exemplified through the current study.

Topic analysis will be further developed by overlaying topics with author communities to explore the interwoven relationships between research topics and research communities (e.g., Yan et al., 2012); by overlaying topics with funding data to investigate the "lead-lag" relationship between funding support and productivity (e.g., Shi et al., 2010); by applying topic models to different genres to study research immediacy (e.g., Ding et al., 2013); and by overlaying topics with citation data to examine the relationships between topics and impact.

References

Åström, F. (2007). Changes in the LIS research front: Time-sliced co-citation analyses of LIS journal articles, 1990–2004. *Journal of the American Society for Information Science and Technology*, *58*(7), 947–957.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3–4), 590–614.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research*, *3*, 993–1022.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., & McAuliffe, J. D. (2010). *Supervised topic models*. arXiv:1003.0783

Bolelli, L., Ertekin, S., Zhou, D., & Giles, C. L. (2009). Finding topic trends in digital libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 69–72). New York: ACM Press.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 155–205.

Chen, C. M. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5303–5310.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386–1409.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. *Physical Review E*, *70*(6), 066111. http://dx.doi.org/10.1103/PhysRevE.70.066111

Ding, Y., Chowdhury, G., & Foo, S. (2000). Incorporating the results of co-word analyses to increase search variety for information retrieval. *Journal of Information Science*, *26*(6), 429–452.

Ding, Y., Yan, E., Sugimoto, C., & Milojevic, S. (2013). Lead-lag topic evolution analysis: Preprints vs. paper. In *Proceedings of the 14th International Conference on Scientometrics and Informetrics (ISSI2013)* July 15–19, Vienna, Austria.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science of the United States of America*, *101*(Suppl. 1), 5228–5235.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Association for Computational Linguistics.

Janssens, F., Zhang, L., Moor, B. D., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology*, *62*(1), 1–18.

Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*, *2001*, 65–72.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology*, *57*(12), 1616–1628.

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., et al. (2010). Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* October 26–30, 2010, Toronto, Canada, (pp. 1565–1568).

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology*, *62*(10), 1933–1953.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5), 056103.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 248–256). Association for Computational Linguistics.

Shi, X., Nallapati, R., Leskovec, J., McFarland, D., & Jurafsky, D. (2010). Who leads whom: Topical lead-lag analysis across corpora. In *NIPS Workshop on Computational Social Science and Wisdom of Crowds*.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science (LIS) dissertations using Latent Dirichlet Allocation (LDA). *Journal of the American Society for Information Science & Technology*, *62*(1), 185–204.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). New York: ACM Press.

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics*, *83*(1), 15–38.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 424–433). New York: ACM Press.

White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science*, *54*(5), 423–434.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327–355.
Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107–2118.
Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology*, *62*(8), 1498–1514.
Yan, E., Ding, Y., Milojevic, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, *6*(1), 140–153.
Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185–193.
Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.