Topics in dynamic research communities: An exploratory study for the field of information retrieval

Erjia Yan[1], **Ying Ding, Staša Milojević, Cassidy R. Sugimoto**
School of Library and Information Science, Indiana University, Bloomington, USA

Abstract

Research topics and research communities are not disconnected from each other: communities and topics are interwoven and co-evolving. Yet, scientometric evaluations of topics and communities have been conducted independently and synchronically, with researchers often relying on homogeneous unit of analysis, such as authors, journals, institutions, or topics. Therefore, new methods are warranted that examine the dynamic relationship between topics and communities. This paper examines how research topics are mixed and matched in evolving research communities by using a hybrid approach which integrates both topic identification and community detection techniques. Using a data set on information retrieval (IR) publications, two layers of enriched information are constructed and contrasted: one is the communities detected through the topology of coauthorship network and the other is the topics of the communities detected through the topic model. We find evidence to support the assumption that IR communities and topics are interwoven and co-evolving, and topics can be used to understand the dynamics of community structures. We recommend the use of the hybrid approach to study the dynamic interactions of topics and communities.

Keywords: community; knowledge discovery; coauthorship; network; Latent Dirichlet Allocation

Introduction

The production of scientific knowledge has become increasingly interdisciplinary and dynamic—geographic, disciplinary, and social boundaries that once isolated scholars are becoming more permeable. In particular, scholars are increasingly mobilized from disparate communities to solve particular problems. This combination and mutual engagement among previously unrelated topic areas benefits both scholars and scholarship (Rodriguez & Pepe, 2010). Noting this change in scientific production, scientists and policy makers have sought better tools for identifying emergent trends and the development of new scholarly communities. However, the classifications for scholarship (e.g., JCR categories, Library of Congress classification) are often inflexible and defective in identifying emerging research fronts and topic bursts (Van Eck & Waltman, 2010). The indices and repositories in which the scholarship is organized, however, provide rich data sources for analyses of cognitive and social developments in the field.

[1] *Correspondence to*: Erjia Yan, School of Library and Information Science, Indiana University, 1320, E. 10th St., LI011, Bloomington, Indiana, 47405, USA. Email: eyan@indiana.edu

To address this, scholars have often examined scholarship using homogeneous variables—examining the growth of new topics using topic analysis techniques or demonstrating the growth of "invisible colleges" through co-citation or collaboration networks. These each provide a single lens on the production of new knowledge—demonstrating novel topics and emergent communities independently. However, research topics and research communities are not disconnected from each other. Communities and topics are interwoven and co-evolving. Therefore, we are motivated to explore how communities interact with topics and how topics co-evolve with communities.

The complexity of scholarly data has led to a growing interest in applying probabilistic models to identify topics from documents. A topic represents an underlying semantic theme and can be informally approximated as an organization of words and can be formally operationalized as a probability distribution over terms in a vocabulary (Blei, 2007). The identification of topics follows the assumption that the more words the two entities share, the more similar these two entities are (Ding, forthcoming). Topic models are the latest advancement in this vein of research (e.g. Blei, Ng, & Jordan, 2003). Topic models provide useful descriptive statistics for a collection of scholarly data, thus making it easier for scholars to navigate academic documents. The outcomes of topic models are probability distributions of words or publications for each topic (e.g. Blei, Ng, & Jordan, 2003); however, they provide no information on which community contributes to a certain topic or how topics are developed by communities. Even though some advanced topic models can generate an author probability distribution for each topic (e.g. Tang et al., 2008), authors belonging to each topic may not necessarily belong to the same community.

Research communities can be detected using community detection methods to group actors, such as authors and journals, with the goal of identifying patterns of community interactions. Qualitatively, a community is a group of associated actors sharing similar characteristics or interests and perceiving or having been perceived as distinctive from the larger society[2]. Since most actors are interacting with each others in certain forms of relations, in scientometrics, a community can be operationalized as a subset of actors densely connected internally and loosely connected externally. Radicchi et al. (2004) gave a quantitative definition of a community: in a strong community each node has more connections within the community than with the rest of the graph ($k_i^{in}(V) > k_i^{out}(V), \forall i \in V$, where k_i is the degree of node i, V is a subgraph). Leskovec et al. (2008) used the concept of "conductance" to capture a community: a good community should have small conductance, i.e. "it should have many internal edges and few edges pointing to the rest of the network" (p. 4). A decisive advance in community detection was made by Newman and Girvan (2004), who introduced a quantitative measure for the quality of partitioned communities, a.k.a. the modularity. In studies of scholarly communications, community detection methods are usually applied to coauthorship networks where the authors are the only nodes, thus leaving us with no information on topics (Ding, 2011). Consequently, one cannot tell in what topic a community is specialized or how communities are related via topics.

Furthermore, topics and communities are not fixed; rather, they develop and evolve dynamically. Some topics are continuously investigated while others appear or disappear over time (Upham &

[2] http://www.merriam-webster.com/dictionary/community

Small, 2010). Similarly, a community may expand or shrink in size, and be divided into several smaller ones or be merged with other communities. Dynamicity is an essential feature of both topics and communities. Studies on topic identification or community detection would be considered as incomprehensive if they fail to capture the dynamic nature of topic or community development.

In reality, communities and topics are not disconnected; on the contrary, communities and topics are interwoven and co-evolving: that is, a research community can carry several topics, and a topic can consist of different collaboration groups (Li et al., 2010). Therefore, in order to study the interdisciplinary nature of science, it is necessary to integrate the two threads of research on community detection and topic identification, and utilize them to understand the dynamic interactions between topics and communities. Questions as the relationship between topics and communities need to be addressed (research questions are formally proposed at the end of the literature review).

An example is used to illustrate the approach of overlaying communities with topics in this study. The upper left image (a) in Figure 1 only contains community information obtained from community detection. As can be seen, authors are partitioned into five clusters but no topic information can be obtained. The upper right image (b) contains 20 topics received from topic models. Obviously, no information can be obtained on which community contributes to which topic. The lower image (c) displays the outcome of adding topics to each community (denoted as C). Besides author partitions, topics (denoted as T) for each community can also be identified. Such approaches allow us to explore the interaction between topics and communities.

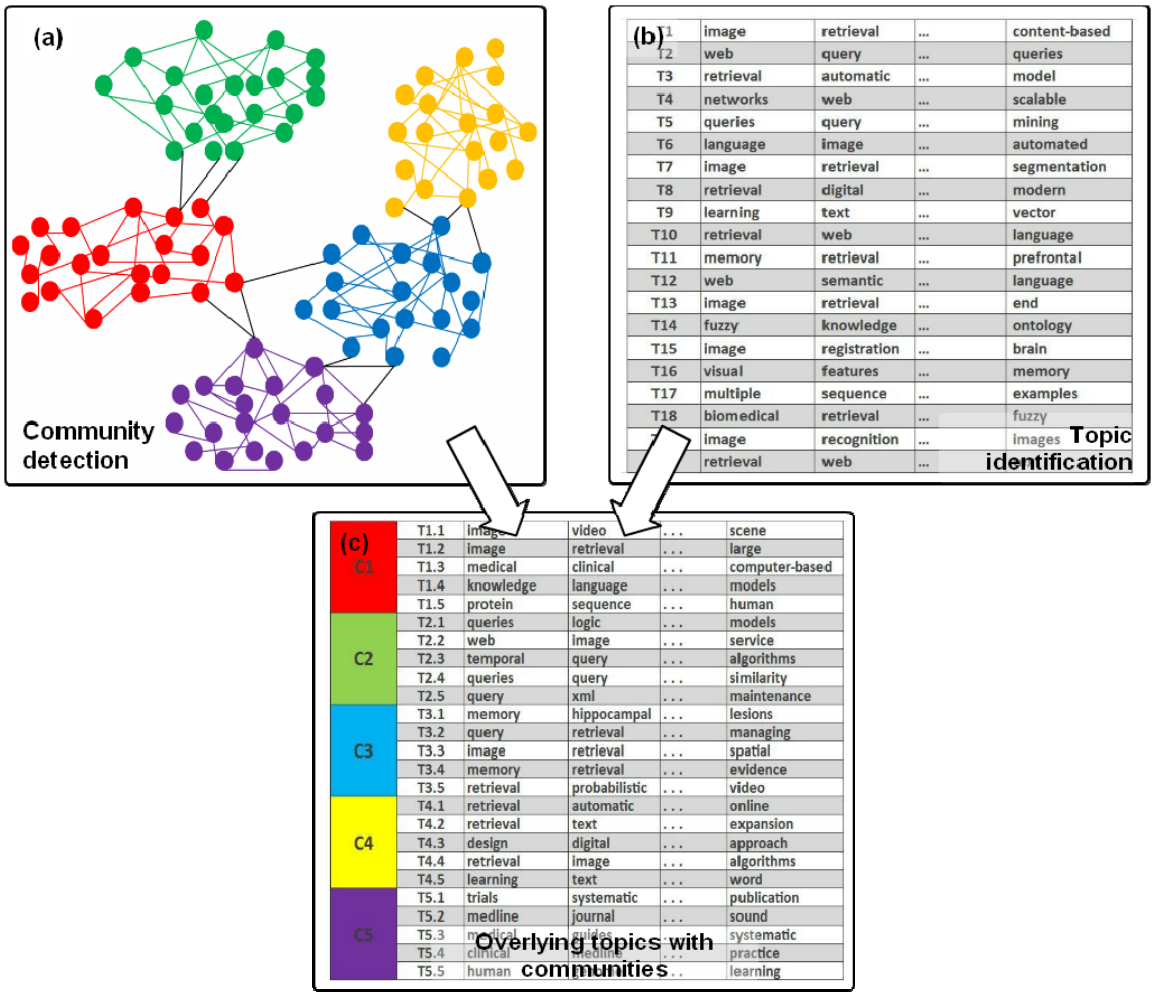

Figure 1. Adding topics to research communities

Information Retrieval (IR) is selected as the test domain. Three coauthorship networks from IR publications are constructed. Research communities are first detected for the three coauthorship networks. Topics are then extracted from IR publications. This study relates communities with topics and extends topic identification to dynamic research communities. The findings of this work contribute to the studies of scholarly communication by exploring how communities interact with topics and how topics co-evolve with communities.

Related work

Detecting author communities

Finding research communities has long been one of the foci of information scientists. The community in bibliometric analysis is represented as clusters of authors, documents, journals, or words. For example, Racherla and Hu (2010) constructed a topic similarity matrix by assigning a predefined research topic to each document and its authors, and using authors' collaboration information to link topics. They found that authors not only collaborate on the same research topics but also collaborate on varied research topics. Upham, Rosenkopf, and Ungar (2010)

developed an iterative clustering scheme that produces high-quality dynamic clusters over time. Using such an approach, twenty-one research communities were detected in the information science and technology area. Innovation performance was then quantified by various parameters and measured for each of these clusters. Pepe and Rodriguez (2010) conducted an in-depth study of a small collaboration network of researchers in the area of sensor networks and wireless technologies. They adopted the notion of discrete assortativity coefficient to evaluate the collaboration pattern in this network. They found that its collaboration has become more intra-institutional and more inter-disciplinary. Giuliani, Petris, and Nico (2010) assessed the collaboration potential for authors in a medical research center. Their assumption is that authors working on similar topics who have not collaborated before are more likely to collaborate in the future. Yet, as Hoekman (2009) pointed out, besides topicality, other factors may also affect collaborations, such as physical, social, and organizational restraints.

Built upon previous endeavors on graph partitioning, Girvan and Newman (2002) proposed an algorithm that uses edge betweenness to identify the boundaries of communities. They applied the method to a scientific collaboration network at the Santa Fe Institute, and identified several densely connected communities. They found that scientists are grouped together either by a similar research topic or by a similar research methodology, where the latter situation may be an indication of interdisciplinary work. Li et al. (2010) constructed a coauthorship network based on authors of the IEEE Transactions on Intelligent Transportation Systems. They applied Girvan-Newman's method to this network and found the collaboration displayed a strong collocation feature where authors of the same institution or the same country are more likely to be coauthors. In addition, they also identified several topics from an author co-word network. Nevertheless, the coauthorship network and co-word network are not systematically integrated, and thus no conclusion is made on the interaction of communities with topics. The Girvan-Newman algorithm is computationally time demanding and is optimized into a more efficient algorithm (Clauset, Newman, & Moore, 2004). The new algorithm incorporated modularity, now becoming a standard measure to evaluate community structures. For instance, Richardson et al. (2009) found their spectral graph-partitioning algorithm can yield higher-modularity partitions. They applied their method to a coauthorship network of network scientists and found three well-known research centers in network science. However, from their findings, it is unclear whether the three locations also form three distinct research topics or how the research centers are connected via topics. The approaches mentioned above can effectively partition nodes into identifiable groups; however, since these networks do not include information on topics, topics cannot be identified from coauthorship network topologies. Consequently, community detection is not able to yield information on what topic a community is specialized in or how communities are related via topics.

Identifying topics

Another thread of research attempting to identify patterns from large scholarly data focuses on detecting topics from documents. Similar to the methods used in detecting author communities, scholars working on identifying topics have used methods such as multidimensional scaling (e.g. White & McCain, 1998), k-means (e.g. Yan, Ding, & Jacob, submitted), modularity-based clustering techniques (e.g. Van Eck & Waltman, 2010), and hybrid approaches (e.g. Janssens,

Glänzel, & De Moor, 2008). Bibliometricians have applied different clustering approaches to identify research fields (e.g. Van Eck & Waltman, 2010), map the backbone of science (e.g. Boyack, Klavans, & Börner, 2005), or portray intellectual landscapes (e.g. Cronin & Meho, 2008) Upham and Small (2010), for instance, gave a good quantitative definition of growing, shrinking, stable, emerging, and exiting research fronts. Traditionally, the research instruments they utilize are mainly co-occurrence networks, for instance, author co-citations networks (White & McCain, 1998), document co-citation networks (Small, 1973; Small, 2006; Klavans & Boyack, 2011; Upham & Small, 2010), journal co-citation networks (Ding, Chowdhury, & Foo, 2000a), or co-word relations (Ding, Chowdhury, & Foo, 2000b; Milojević, Sugimoto, Yan, & Ding, 2011). Currently, there is a trend in bibliometrics of using hybrid approaches to identify topics in scientific fields. Liu et al. (2010) presented a framework of hybrid clustering to combine lexical and citation data for journal set analysis. Their hybrid approach can be employed as a good reference for journal categorization. Zitt, Lelu, and Bassecoulard (2011) examined the convergence of two thematic mapping approaches: citation-based and word-based. They found the two approaches yield quite different outcomes and cannot substitute each other. Boyack and Klavans (2010) examined several types of scholarly networks, including a cocitation network, a bibliographic coupling network, and a citation network, in the interest of selecting the network that can represent the research front in biomedicine. They used within-cluster textual coherence and grant-to-article linkage indexed by MEDLINE as accuracy measurements and found that the bibliographic coupling-based citation-text hybrid approach, an approach that couples both references and words from title/abstract, outperformed other approaches. Janssens, Glänzel, and De Moor (2007, 2008) proposed a novel hybrid approach that integrates two types of information, citation (in the form of a term-by-document matrix) and text (in the form of a cited_references-by-document matrix). Noticing that the weighted linear combinations may "neglect different distributional characteristics of various data sources" (p. 612), the authors developed a new approach named Fisher's inverse chi-square method. This method can effectively combine matrices with different distributional characteristics. They found the hybrid approach outperformed the text-only approaches by successfully assigning papers into correct clusters.

Above mentioned studies on identifying topics yield discrete assignments: a node is usually assigned to one cluster. In this sense, they are closely related to community detection research. There are studies on identifying topics that use topic models and yield fractional assignments (probability distributions). Followed by the tradition of data mining and knowledge discovery, topic models have gained great popularity among computer scientists in recent years. One well-known topic model is the Probabilistic Latent Semantic Indexing (pLSI) model proposed by Hofmann (1999). Built on pLSI, Blei et al. (2003) introduced a three-level Bayesian network, called Latent Dirichlet Allocation (LDA). In topic models, topics are modeled as a probability distribution over terms in a vocabulary. Topic models have also been extended to include authorship information. Steyvers et al. (2004) proposed an unsupervised learning technique for extracting both the topics and authors of documents. In their Author-Topic model, authors are modeled as probability distributions over topics. McCallum et al. (2004) presented the Author-Recipient-Topic (ART) model, a directed graphical model which conditions the per-message topic distribution jointly on both the author and individual recipients. In ART model, each topic is modeled as a multinomial distribution over words, and each author-recipient pair is modeled as a

distribution over topics. The Author-Conference-Topic (ACT) Model, proposed by Tang et al. (2008), further extended Author-Topic model to include conference/journal information. The ACT model utilizes probabilistic models to model documents' contents, authors' interests, and also conference/journal simultaneously. As noted, topic models can only generate an author probability distribution over each topic; yet authors belonging to each topic may not belong to the same community. Hence, topic models still cannot address how communities interact with topics.

Overlaying communities with topics

To understand the interaction between research communities and research topics, there is a need to incorporate both community detection and topic modeling approaches. For instance, Zhou et al. (2006) proposed two generative Bayesian models for semantic community detection in social networks by combining probabilistic modeling with community detection algorithms. Their method was able to detect the communities of individuals and meanwhile provide topic descriptions to these communities. Li et al. (2010) combined LDA with the Girvan-Newman's community detection algorithm and tested their method on a social tagging data set. They found that communities and topics are interwoven and co-evolving. Hybrid approaches can integrate different types of scholarly networks, for example, citation-based and word-based networks (Janssens, Glänzel, & De Moor, 2008; Liu et al., 2010; Zitt, Lelu, & Bassecoulard, 2011) and co-occurrence networks (Boyack & Klavans, 2010), but these studies were largely focused on providing more precise clustering results but did not address the interactive nature of research topics and communities.

To answer this, our study presents a hybrid approach to study the interaction between topics and communities and evaluates this approach on the Information Retrieval domain. Therefore, the question we seek to answer is:

1. What is the relationship between topics and communities in the Information Retrieval domain?

   The current scholarship tends to study topics and communities separately; however, topics and communities are not disconnected: a research community can carry several topics, and a topic can be studied by different collaboration groups. In answering this, we hope to demonstrate how incorporating elements of topic and community can lead to an enhanced understanding of the domain. The mutual engagement of various academic entities (e.g. papers, authors, journals, words, etc.), on the one hand, provides opportunities to scientometricians, as a type of academic entity can now be studied in relation to other entities from multiple perspectives; on the other hand, it brings challenges as well, as the complexity increases significantly when more heterogeneities are added to scholarly networks. Therefore, in order to discover patterns from the complexities, it is necessary to understand the relationship between different academic entities. The present research addresses this issue by studying the relationship between topic and community in the Information Retrieval domain.

2. How can this hybrid approach be used to enhance our understanding of the dynamic interaction between topics and communities of a domain?

By incorporating both topic identification and community detection approaches, we are able to obtain a more holistic understanding of the dynamicity of a domain. Furthermore, this paper presents a novel methodological approach; therefore, one of the research objectives is to explicate the process, provide examples of appropriate visualization techniques, and demonstrate the value of such a hybrid approach.

Data on IR were chosen to exemplify our approach. The ACT model was selected as it is a recent advance in topic models; Clauset-Newman-Moore method was selected as it is the most used and best known community detection method (Fortunato, 2010).

Methods

Data

Information retrieval (IR) was chosen as the target domain. Papers were collected from Scopus for 2001-2007 (inclusive). Coauthorship networks were constructed based on all authors. Author name disambiguation is a complicated task. Ideally, each name stands for a unique author; however, two types of errors may be generated: different names may attribute to the same author (e.g. Jacob, E.K. and Jacob, E. may both refer to the same author Elin K. Jacob), and a name may be attributed to a single author when it represents multiple—a common error with Asian names (e.g. Wang, L. may be the name of several authors). Radicchi et al. (2009) merged LAST-NAME, F. M. and LAST-NAME, FIRST-NAME MIDDLE-NAME into same author. Yan and Ding (2009) combined the same authors manually based on their affiliation information. Milojevic (2009) compared the slopes of degree distributions of using all initials and using first initials, and found using fist initials had more precise match with power-law distribution. Barabasi et al. (2002), however, argued that for coauthorship networks, author disambiguation may not be critical. Moody (2004) found no significant difference in the results in coauthorship networks using the methods for name disambiguation.

Author names were processed by identifying outliers through publication frequency, a practical method proposed by Newman (2001). One hundred and fifteen authors who have published more than eight papers per year were identified. Google Scholar (in Engineering, Computer Science, and Mathematics) and DBLP were used to verify whether the high quantity is the result of productivity or the result of repetitive names. Only eight authors out of 115 were actual individual authors who were productive, while the rest were attributed to repetitive names. In order to minimize the negative influence of repetitive names, records of the repetitive names were deleted—these deleted records represent approximately 2% of the total records.

Time slices were set as 2001-2003, 2004-2005, and 2006-2007 so that each slice has similar number of authors, thus providing comparable networks. Authors in the largest component (LC) were finally selected to form the coauthorship networks.

Table 1. Data statistics

	2001-2003	2004-2005	2006-2007
No. of papers	12,194	19,145	21,423
No. of authors in the LC	7,354	14,213	17,710

Approaches

Detecting research communities

Clauset, Newman, and Moore's (2004) method was implemented to the coauthorship networks for each time period. The modularity for weighted networks can be calculated as (Clauset, Newman, & Moore, 2004):

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (1)$$

A_{ij} is the weight of the connection from i to j; m denotes the total number of links in the network, which is $\frac{1}{2}\sum_{ij} A_{ij}$; k_i is the degree of a vertex i in a weighted network, which is $\sum_j A_{ij}$; δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise.

Formula (1) is the fraction of within-community edges $\frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i, c_j)$ minus the expected value of the same degrees of vertices randomly connected between the vertices: $k_i k_j/2m$.

Detecting research topics

An extended stop word list is used to exclude common words in IR, including information, retrieval, system, search, and model[3]. The ACT model (Tang et al., 2008) was used to detect topics. In the ACT model, each author is associated with a multinomial distribution over topics and words in a paper and the conference stamp is generated from a sampled topic. The generative process of the ACT model can be summarized as follows:

1. For each topic z, draw ϕ_z and ψ_z respectively from Dirichlet priors β and μ (ϕ_z :the multinomial distribution over words specific to z; ψ_z : the multinomial distribution of publication venues specific to topic z);
2. For each word w_{di} in paper d:
   - draw an author x_{di} from a_d uniformly (a_d : vector form of authors in paper d);
   - draw a topic z_{di} from a multinomial distribution $\theta_{x_{di}}$ specific to the author x_{di}, where θ is generated from a Dirichlet prior α ;
   - draw the word w_{di} from multinomial $\phi_{x_{di}}$; and
   - draw the conference stamp c_{di} from multinomial $\psi_{x_{di}}$.

In this way, the posterior distribution of topics depends on three modalities: authors, words, and conferences (or journals). The model begins with the joint probability of the whole data set, and

[3] http://ella.slis.indiana.edu/~eyan/papers/stoplist.txt

then using the chain rule, the posterior probability of sampling the topic and author for each word can be obtained. Then by using the chain rule, the posterior probability of sampling the topic z_{di} and the author x_{di} for the word w_{di} is:

$$P(z_{di}, x_{di} \mid \mathrm{z}_{-\mathrm{di}}, \mathrm{x}_{-\mathrm{di}}, \mathrm{w}, \mathrm{c}, \alpha, \beta, \mu) \propto \frac{P(\mathrm{z}, \mathrm{x}, \mathrm{w}, \mathrm{c}, \alpha, \beta, \mu)}{\mathrm{P}(\mathrm{z}_{-\mathrm{di}}, \mathrm{x}_{-\mathrm{di}}, \mathrm{w}, \mathrm{c}, \alpha, \beta, \mu)}$$
$$\propto \frac{m^{-di}_{x_{di} z_{di}} + \alpha}{\sum_z (m^{-di}_{x_{di} z} + \alpha)} \frac{n^{-di}_{x_{di} w_{di}} + \beta}{\sum_v (n^{-di}_{z_{di} v} + \beta)} \frac{n^{-d}_{z_{di} c_d} + \mu}{\sum_c (n^{-d}_{z_{di} c} + \mu)} \quad (2)$$

where m_{xz} is the number of times that topic z has been used associated with author x, n_{zv} is the number of times that word w_v has been generated by topic z, n_{zcd} is the number of times that conference c_d generated by topic z. z_{-di} and x_{-di} represent all topics and authors assignments excluding the i-th word in the paper d; the numbers m^{-di} and n^{-di} with the superscript $-di$ denote a quantity, excluding the current instance (the i-th word token or the conference stamp in the paper d). Since the estimated topic models are not very sensitive to the hyperparameters, for simplicity, they were set as fixed values (i.e., $\alpha = 50/\mathrm{T}$, $\beta = 0.01$, and $\mu = 0.1$).

Overlaying topics with communities

The next step was to overlay research topics for the detected communities. The procedures were: (1) search and collect publications for all authors in the top ten communities in each time slice; (2) apply the ACT model to publications of each time slice with the number of topics set at ten; (3) generate an topic-author distribution (P(topic | author)) using the ACT model where each author obtains a topic distribution vector (for author i: $a_i = (t_1, t_2, ..., t_{10})$), and set up a threshold and replace those probabilities that below the average 0.1 (1/10) to 0; by doing so, the insignificant probabilities will not be counted and will not add noise to the community similarity calculation; (4) extract and average the topic distributions for authors of a community where the mean is considered as the community's topic distribution vector, and then normalize the vector so that the sum of each vector is one; (5) calculate cosine similarities for communities. An example is used to illustrate the last three steps. There are ten authors (A1, A2,…, A10) in Figure 2. They belong to three communities (C1, C2, and C3).

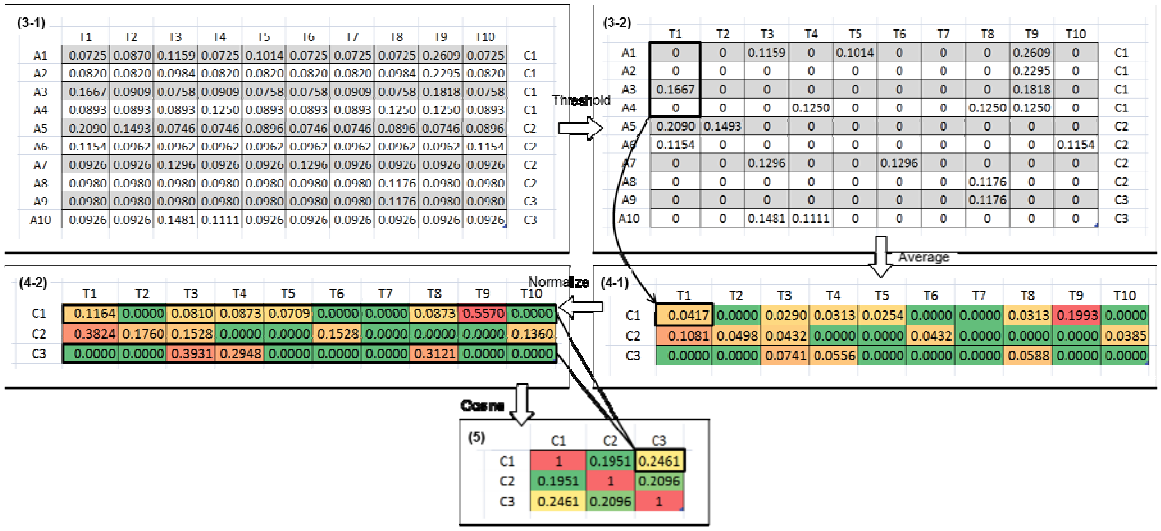

Figure 2. An example of overlaying topics with communities

Results and analyses

Detecting research communities

This section first examines the author dynamics in the LC, and then delves into the author dynamics at the community level. Figure 3 illustrates the adoption of authors in the LC from 2001-2003 to 2006-2007.

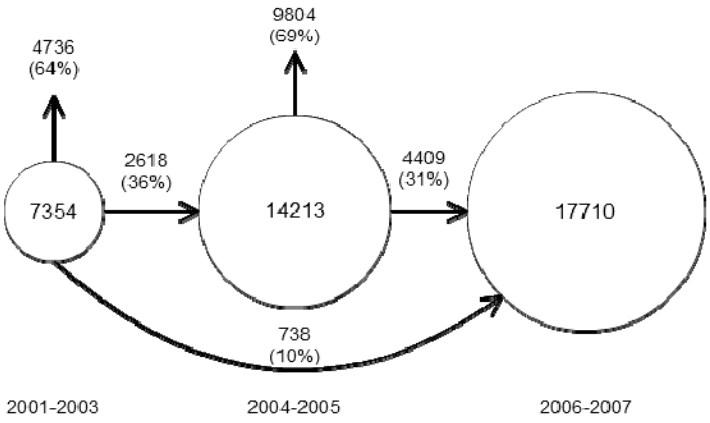

Figure 3. Author dynamics

The number of authors in the LC increased from 7,354 in 2001-2003 to 17,710 in 2006-2007, indicating that more scholars have joined the IR research community. More than 60% of the authors who previously published papers on IR no longer published papers in this field, and they were replaced by new scholars. At the same time, around 30% of the authors continuously published papers on IR. In addition, 10% of the authors in 2001-2003 skipped the 2004-2005 period and resumed publishing papers in 2006-2007.

Table 2 shows the sizes of top ten communities for the three time periods.

Table 2. Size of communities

	2001-2003	2004-2005	2006-2007
Number of communities	82	246	293
Size of largest community	728	3026	3736
Size of second community	487	1277	1330
Size of third community	333	622	862
Size of fourth community	272	546	644
Size of fifth community	224	412	633
Size of sixth community	223	320	574
Size of seventh community	221	307	540
Size of eighth community	209	289	434
Size of ninth community	194	287	386
Size of tenth community	191	229	339
Ratio of top ten communities	41.91%	51.47%	53.52%

The top ten communities have an extensive coverage as they represent around half of the authors in the LC. Therefore, in the following paragraphs, the top ten communities are used as the unit of analysis.

Dunbar (1998) predicted that 150 is roughly the upper limit of a well-functioning human community. Several other studies also found that smaller communities are desirable, for example Allen (2004) found that on-line communities usually have 60 members, and if there are more than 80 members the community will break down and end up in several smaller new communities. Leskovec et al. (2008) found that communities of size beyond 100 nodes gradually blend into the core of the network and thus become less community-like "with a roughly inverse relationship between community size and optimal community quality" (p. 1). As a link in coauthorship networks is merely a proximation of collaboration relationship in real life, it may not be direct collaboration: scholars may appear as coauthors in an article but they may not necessarily maintain collaboration relationships in their academic life. As a result, the sizes of the clusters in coauthorship networks are usually larger than the Dunbar's number 150. It also indicates the need to conduct bibliometric studies on a more focused and scalable size.

Table 3 and Table 4 match authors in consecutive time periods among the top ten communities. For example, the number 143 means that 143 authors in the largest community in 2004-2005 are coming from the largest community in 2001-2003.

Table 3. Matching between 2001-2003 and 2004-2005 communities

		2004-2005 communities									
		1	2	3	4	5	6	7	8	9	10
2001-2003 communities	1	**143**	**39**	6	0	1	8	3	9	9	0
	2	126	11	5	1	3	4	0	1	5	0
	3	36	20	3	3	2	2	2	2	2	0
	4	4	0	**35**	1	1	1	3	0	5	0
	5	15	14	2	**16**	9	0	0	0	1	0
	6	31	8	0	0	0	0	0	0	**15**	0
	7	12	1	3	2	**19**	1	1	3	4	1
	8	46	3	0	0	0	1	0	4	2	0
	9	12	3	3	2	4	4	1	4	3	1
	10	11	9	0	1	1	**20**	0	0	5	1

	sum(1-10)	436	108	57	26	40	41	10	23	51	3
	sum(rest)	298	163	71	97	26	16	48	49	19	20
	new	2292	1006	494	423	346	263	249	217	217	206
	total	3026	1277	622	546	412	320	307	289	287	229

The highest overlapping for each community is displayed in bold. Notably, more than half of the authors in each of the top ten communities in 2004-2005 are new authors. The results suggest that the research communities in IR are expanding, but unstable. New collaborations were formed among new authors; meanwhile some of the existing collaborations were not maintained, meaning that community structures in 2001-2003 were not kept in 2004-2005.

Table 4. Matching between 2004-2005 and 2006-2007 communities

		2006-2007 communities									
		1	2	3	4	5	6	7	8	9	10
2004-2005 communities	1	**654**	44	**30**	**51**	**63**	12	16	**31**	22	3
	2	84	**204**	26	12	3	5	9	11	2	1
	3	32	20	8	3	2	2	**53**	5	1	5
	4	21	1	6	27	1	**61**	2	3	0	0
	5	23	2	14	1	18	17	3	6	**27**	0
	6	25	14	10	1	9	3	2	5	2	0
	7	14	9	14	1	2	1	17	6	1	6
	8	47	12	8	3	3	3	2	12	3	0
	9	32	7	8	6	10	2	3	5	4	7
	10	13	4	3	2	1	1	4	4	0	0
	sum(1-10)	945	317	127	107	112	107	111	88	62	22
	sum(rest)	373	139	125	90	57	50	57	28	35	9
	new	2418	874	610	447	464	417	372	318	289	308
	total	3736	1330	862	644	633	574	540	434	386	339

Similar to preceding analysis, the majority of authors in the 2006-2007 communities are new authors. What differs Table 4 from Table 3 is that in Table 4 around 20% of authors in the largest two communities in 2006-2007 are coming from the same communities in 2004-2005, indicating that communities are stabilizing in recent time periods. Note that this may also be the result of the increased cluster sizes, as in the latter two time periods communities are larger which may lead to higher likelihood of these communities containing common authors.

Detecting research topics

This section examines the word dynamics of all publications by the authors in the largest ten communities, reports the topic popularity obtained from the ACT model, presents the heat map based on the topical cosine similarity matrix, and uses correspondence graph to illustrate how topics are semantically connected. Figure 4 illustrates the adoption of title words from 2001-2003 to 2006-2007.

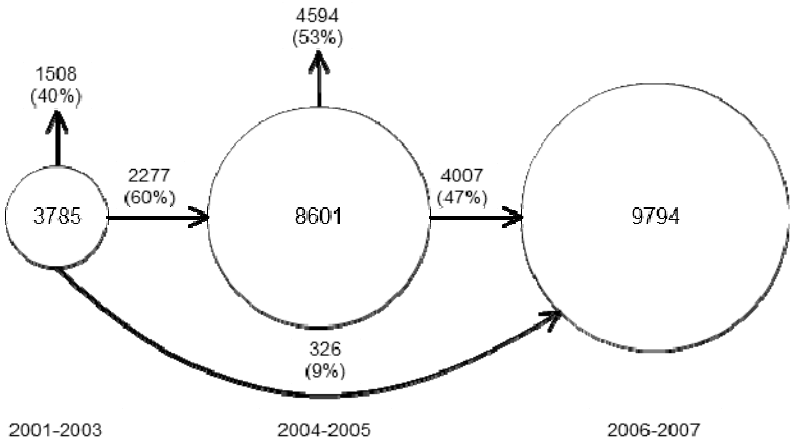

Figure 4. Word dynamics

An increasing number of words have been added to the knowledge domain of IR over time: from 3,785 in 2001-2003 to 9,794 in 2006-2007, indicating an expanded research scope of IR scholars. Around half of the words used in the earlier period are inherited by the next period—the other half is abandoned. In addition, 10% of the words in 2001-2003 were not mentioned in 2004-2005 but regained attention in 2006-2007.

Topic popularity is predicated on the ACT model. The underlying assumption is that if the words belonging to a certain topic occur more frequently, then this topic has high popularity. Since ten topics are set, a topic popularity of 0.1 means this topic has an average popularity. A value above 0.1 suggests a "hot" topic and a value below 0.1 suggests a "cold" topic. In Table 5, topics for each period are ranked based on topic popularities, and for remaining analysis, the same rank is followed (the labels for each topic can be found in Figure 6).

Table 5. Topic popularity

	2001-2003	2004-2005	2006-2007
Topic 1	0.1250	0.1160	0.1227
Topic 2	0.1201	0.1152	0.1141
Topic 3	0.1193	0.1135	0.1127
Topic 4	0.1118	0.1067	0.1104
Topic 5	0.0972	0.1062	0.1095
Topic 6	0.0966	0.0994	0.1016
Topic 7	0.0901	0.0978	0.0878
Topic 8	0.0860	0.0944	0.0863
Topic 9	0.0799	0.0792	0.0839
Topic 10	0.0740	0.0716	0.0708

For topics from the same time period, the calculation of topic similarities can be made directly as they share the same array of words. However, for topics from different time periods, extra steps are needed to calculate topic similarities. First, the union of all unique words in the three time periods is identified; then, for those words that did not show up in certain time period, their word-

topic distribution (P(word | topic)) is filled with zeros. Therefore, topics from different time periods contain the same array of words, for topic i: $t_i = (w_1, w_2, ..., w_n)$. Cosine similarity is finally calculated for every pair of word-topic distributions. A heat map visualization is shown in Figure 5.

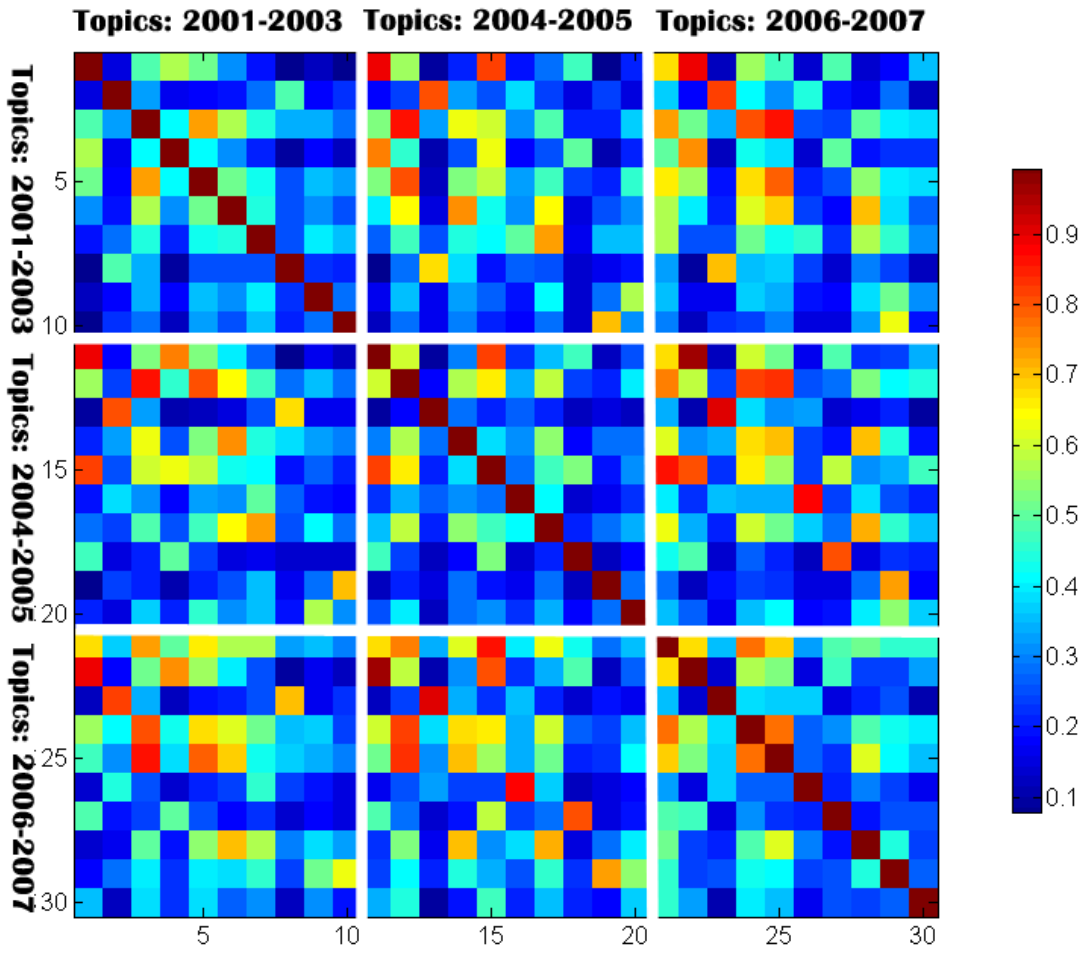

Figure 5. Heat map visualization of topic similarities

For topics belonging to the same time period (the three blocks located on the diagonal line), most topics have low similarities with other topics. It is a good sign in that the ACT model has successfully identified distinguishable topics. For topics belonging to different time periods, it can be found that some topics have evident successors (bright squares) while other topics fail to proceed into the next time period (dark squares). In addition, it can also be found that topics with high popularities tend to have multiple successors and topics with low popularities tend to have only one or none successor. For example, Topic 1 in 2004-2005 has two evident successors: Topic 1 and Topic 2 in 2006-2007; Topic 2 in 2004-2005 has three evident successors: Topic 1, Topic 4, and Topic 5; on the other hand, Topic 9 in 2004-2005 only has one successor: Topic 9 in 2006-2007; and Topic 10 in 2004-2005 does not have identifiable successors.

In order to provide a more informative presentation of what these topics are, Figure 6 is introduced where for each topic, the top five words based on word-topic distribution (P(word | topic)) are listed.

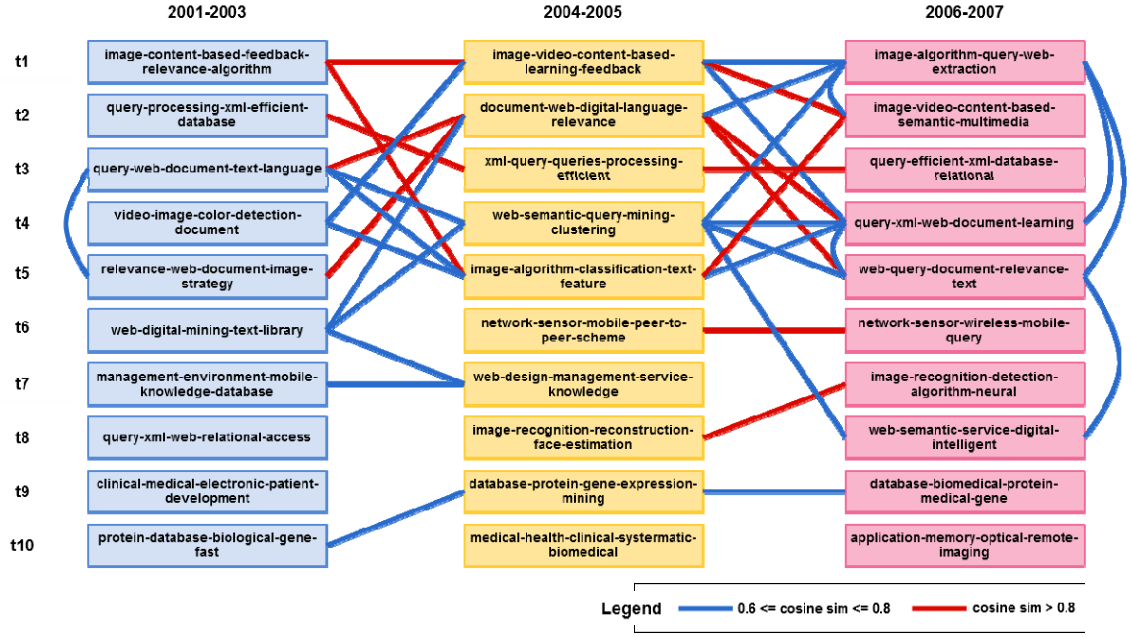

Figure 6. Topic dynamics (blue line: cosine similarity [0.6, 0.8]; red line: cosine similarity (0.8, 1])

Topics in high popularities are well connected: the top five topics in each time period have predictors and/or successors. However, topics in low popularities are loosely connected, suggesting that they did not receive continuous attention. Two types of topics can be identified, including continuous topics and rising topics. Continuous topics denote those topics that are continuously linked though 2001-2003 to 2004-2005 and through 2004-2005 to 2006-2007. Those topics received continuous attention in the past decade, such as "image-algorithm-query-web-extraction" and "image-video-content-based-semantic-multimedia". Rising topics denote those topics that gained attention in later two periods, such as "image-recognition-detection-algorithm-neural" and "application-memory-optical-remote-imaging". Noticeably, biomedical and web application related topics became more popular in recent time periods.

Overlaying topics with communities

The heat map visualization of overlaying topics with communities is illustrated in Figure 7. Figure 7 visualizes community × topic matrices and thus can be read from two directions: each row shows, for each community, how many evident topics this community is specialized in; each column shows that for each topic, how many evident communities are working on it.

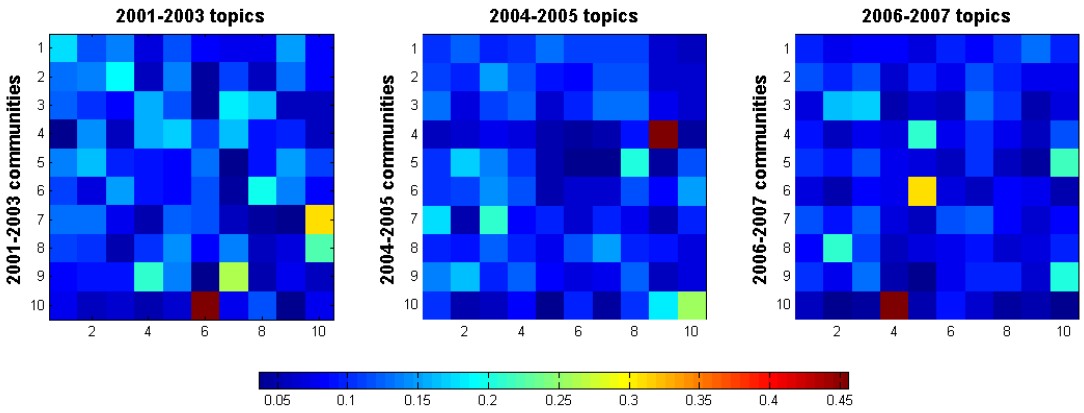

Figure 7. Heat map visualization of community × topic matrices

Communities of smaller sizes tend to have evident topical concentrations, which is understandable as communities of larger sizes are more likely to involve scholars with diverse research interests. For example, in 2001-2003, Community 7 is specialized in Topic 10, Community 9 is specialized in Topic 7, and Community 10 is specialized in Topic 6; comparatively, the top three communities in 2004-2005 and 2006-2007 did not yield evident topical concentrations. In regard to topics, most topics are associated with at least one community. For example, Topic 9 in 2004-2005 is studied by Community 4, and Topic 4 in 2006-2007 is studied by Community 10. The results indicate that authors are more inclined to collaborate with others who have similar expertise and publish papers on similar topics. In addition, smaller communities tend to have relatively distinct research topics. Figure 8 associates communities with their specialized topics.

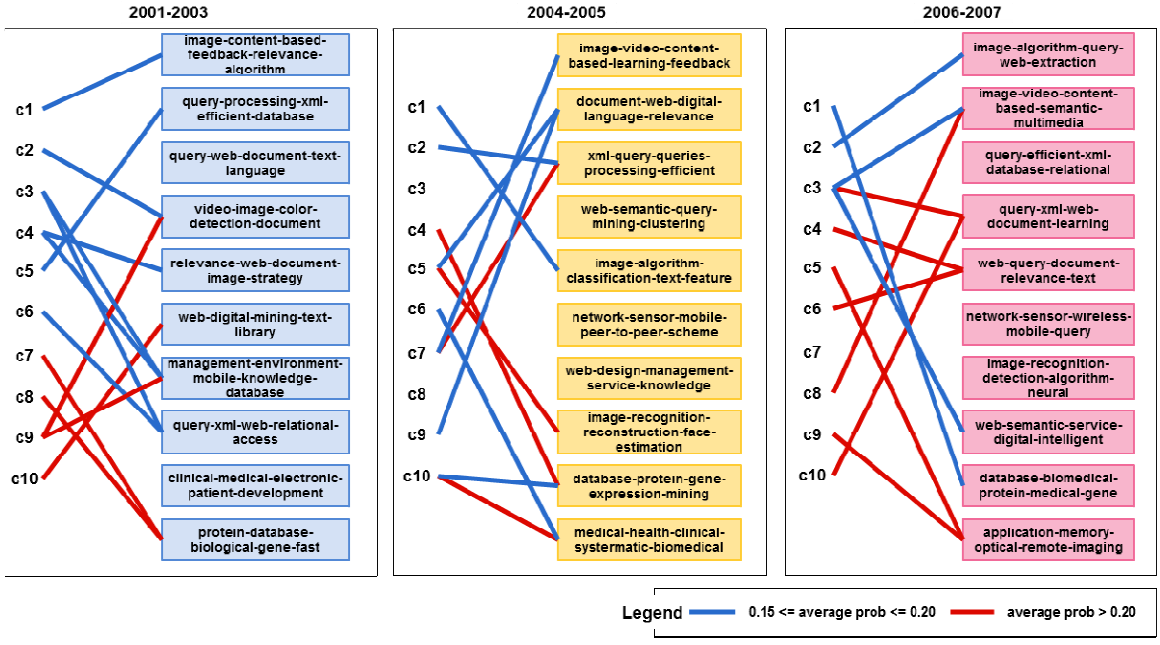

Figure 8. Association between communities and topics (blue line: average probability [0.15, 0.20]; red line: average probability (0.20, 1])

Similar to the above analysis, communities of smaller sizes are specialized on more distinct topics (average probability larger than 0.2). However, if both weak and strong associations are considered (average probability larger than 0.15), it can be found that topics with higher popularities tend to be studied by a greater number of authors. For instance, Topic 1 in 2001-2003 is studied by Community 1; Topic 1 in 2006-2007 is studied by Community 2; Topic 2 in 2006-2007 is studied by Community 3. There are several topics with no discernable community, such as "network-sensor-mobile-peer-to-peer-scheme" in 2004-2005 and "network-sensor-wireless-mobile-query" in 2006-2007. We argue that authors from different communities may contribute to these new topics at beginning as they are emerging research topics. These authors may eventually collaborate with each other more frequently and form a community of their own as they mature.

In Figure 7, each community has a topic distribution, for community i: $c_i = (t_1, t_2, ..., t_{10})$. Based on such distributions, community topical similarity can be obtained through calculating the cosine similarities. The heat map visualization is displayed in Figure 9.

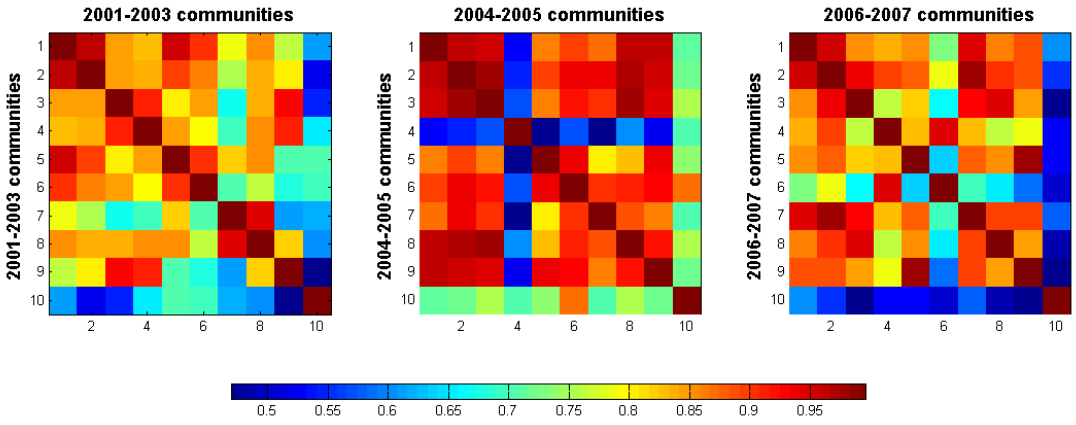

Figure 9. Heat map visualization of community topical similarities

A few communities (Community 7, 9, and 10 in 2001-2003; Community 4 and 10 in 2005-2006; Community 6 and 10 in 2006-2007) concentrates on relatively unique topics and has lower level of topical similarity with other communities, especially for biomedical related topics, such as Community 4 in 2004-2005 is highly specialized in "database-protein-gene-expression-mining", Community 10 in 2004-2005 is highly specialized in "medical-health-clinic-systematic-biomedical", and Community 10 in 2006-2007 "application-memory-optical-remote-imaging". The rest communities have higher level of topical similarity with each other. Within one community, authors may not share homogenous research topics but have several topics which may relate to the topics of other communities. We argue that in IR authors not only collaborate with others who share similar research specialty but also collaborate with scholars from varied domains to enhance their research capability. This is especially evident for application driven research topics which are heavily dependent on labs. The rise of large-scale data collection efforts also generates a similar team-production model (Moody, 2004) where team members usually have different specialties.

Conclusion and Future Research

In this study, a hybrid approach was proposed which integrates both topic identification and community detection techniques. Two layers of enriched information are constructed based on the bibliographic data: one is the topology of the coauthorship network and the other is the topic model of the communities overlaid on the coauthorship network. This work applied the hybrid approach to the domain of Information Retrieval (IR) as a proof-of-concept exercise. We used this case study to confirm the benefit of using the hybrid approach - that is the hybrid approach can lead to an enhanced understanding of a domain. The proposed approach effectively finds evidence to support the interactive nature of topics and communities. The findings provide a novel description of the developments in IR, and also provide a foundation for future research using hybrid approaches.

The study demonstrated that, between 2001 and 2007, only 30% of the authors continuously published in the field. This may imply instability in the field, or a high degree of permeability. Permeability has been used to describe application-focused domains (Klein, 1996), many of which have high technology-dependence. However, as an exploratory study, there is no indication as to how this compares to other domains. Areas of future research should explore the degree of stability in communities to establish baselines for comparison. Similarly, the study showed that the top ten communities represent about half of the total authors in the largest component. Comparisons with other domains will provide an indication with the degree to which this shows high or low community coherence.

By incorporating both topic identification and community detection approaches, we are able to obtain a more holistic understanding of the dynamicity of a domain. The dynamicity in communities supports the need for studies that evaluate scientific developments diachronically. In addition, the large influx of new words over this short time period reinforces the need to study topic development in short intervals and diachronically.

The proposed hybrid approach also provides a lens on topic development—providing an initial exploration of the way in which topics emerge and the popularity factors that sustain a new topic. For the domain of IR, topics of higher popularities tended to be further studied in the succeeding time periods; yet topics of lower popularities received less attention and even vanished from the research focus. Biomedical and web application related topics are becoming more popular in recent time periods.

The results provide an example of the inter-relationship between topics and communities. Our approach shows that in IR, topics are sustained by the creation of a community around these topics; communities are, to a large degree, enhanced by these new topic areas. The approach illustrates the importance of studying the development of science from both cognitive and social perspectives, as the dynamic changes of community structures can contribute to the scholarly communications and can also be used to predict future interactions or shift of topics. The future community detection methods will be focused on the dynamic changes of communities and topics and figure out how important scholars move topics forward.

References

Allen, C. (2004). Life with alacrity: The Dunbar number as a limit to group sizes. Retrieved April 19, 2010 from http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html

Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and 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. *Journal of Machine Learning Research, 3*, 993-1033.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. *Physical Review E, 70*, 066111

Cronin, B., & Meho, L. I. (2008). The shifting balance of intellectual trade in information studies. *Journal of the American Society for Information Science & Technology, 59*(4), 551-564.

Ding, Y. (forthcoming). Community Detection: Topological vs. Topical. *Journal of Informetrics*.

Ding, Y., Chowdhury, G., & Foo, S. (2000a). Journal as markers of intellectual space: Journal co-citation analysis of information retrieval area, 1987-1997. *Scientometrics, 47*(1), 55-73.

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

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Boston, MA: Harvard University Press.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3-5), 75-174.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821–7826.

Giuliani, F., Petris, M., & Nico, G. (2010). Assessing scientific collaboration through coauthorship and content sharing. *Scientometrics, 85*(1), 13-28.

Hoekman, J., Frenken, K., & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *Annals of Regional Science, 43*, 721-738.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 50-57, Aug 15-19, 1999, Berkeley, CA, USA.

Janssens, F., Glänzel, W., & De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.360-369, August 12-15, 2007, San Jose, California, USA.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607–631.

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.

Klein, J.T. (1996). *Crossing boundaries: Knowledge, disciplinarities, and interdisciplinarities*. Charlottesville, VA: University Press of Virginia.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Retrieved March 21, 2010 from http://arxiv.org/abs/0810.1355

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., & Li, J. (2010). Community-based topic modeling for social tagging. The 19th ACM International Conference on *Information and Knowledge Management (CIKM2010)*, pp: 1565-1568, Oct 26-30, 2010, Toronto, Canada

Li, L., Li, X., Cheng, C., Chen, C., Ke, G., Zeng, D., & Scherer, W. T. (2010). Research collaboration and ITS topic evolution: 10 years at T-ITS. *IEEE Transactions on Intelligent Transportation Systems, 11*(3), 517 - 523.

Liu, X., Yu, S., Janssens, F., Glanzel, W., Moreau, Y., & de Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology, 61*(6), 1105-1119.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2004) The Author-Recipient-Topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical Report UM-CS-2004-096. Retrieved May 30, 2010 from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.5833

Milojevic, S. (2009). *Big science, nano science? : Mapping the evolution and socio-cognitive structure of nanoscience/nanotechnology using mixed methods*. Doctoral dissertation, University of California, Los Angeles.

Milojevic, 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.

Moody, J. (2004). The Structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*, 213-238.

Newman, M. E. J. (2001). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E, 64*, 016131.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

Pepe, A., & Rodriguez, M. A. (2010). Collaboration in sensor network research: an in-depth longitudinal analysis of assortative mixing patterns. *Scientometrics, 84*(3), 687-701.

Racherla, P., Hu, C. (2010). A social network perspective of tourism research collaborations. *Annals of Tourism Research, 37*(4), 1012-1034.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2658-2663.

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

Richardson, T., Mucha, P. J., & Porter, M. A. (2009). Spectral tripartitioning of networks. *Physical Review E, 80*, 036111.

Rodriguez, M. A., & Pepe, A. (2010). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics, 2*(3), 195-201.

Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68(3), 595-610.

Small, H.G. (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*, 265-269.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306-315, New York: ACM Press.

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.

Upham, S. P., Rosenkopf, L., & Ungar, L. H. (2010). Innovating knowledge communities: An analysis of group collaboration and competition in science and technology. *Scientometrics, 83*(2), 525-554.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

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 and Technology, 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., Ding, Y., & Jacob, E. (forthcoming). Overlaying communities and topics: An analysis on publication networks. *Scientometrics*.

Zhou, D., Ji, X., Zha, H., & Lee Giles, C. (2006). Topic evolution and social interactions: How authors affect research. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, November 6-11, 2006, Arlington, Virginia, USA.

Zhou, D., Manavoglu, E., Li, J., Lee Giles, C., & Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th ACM International Conference on World Wide Web*, May 23-26, 2006, Edinburgh, Scotland.

Zitt, M., Lelu, A., & Bassecoulard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 19-39.