Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval

Baitong Chen
Department of Library, Information and Archives, Shanghai University, Shanghai 200444 CHINA.
School of Information Management, Wuhan University, Wuhan 430072 CHINA.
E-mail: baitongchen@shu.edu.cn

Satoshi Tsutsui
School of Informatics and Computing, Indiana University, IN 47408 USA. E-mail: stsutsui@indiana.edu

Ying Ding
School of Informatics and Computing, Indiana University, IN 47408 USA.
School of Information Management, Wuhan University, Wuhan 430072 CHINA.
University Library, Tongji University, Shanghai 200092 CHINA
Email: dingying@indiana.edu

Feicheng Ma
School of Information Management, Wuhan University, Wuhan 430072 CHINA. E-mail: fchma@whu.edu.cn

Abstract

Understanding topic evolution in a scientific domain is essential for capturing key domain developments and facilitating knowledge transfer within and across domains. Using a data set on information retrieval (IR) publications, this paper examines how research topics evolve by analyzing the topic trends, evolving dynamics, and semantic word shifts in the IR domain. Knowledge transfer between topics and the developing status of the major topics have been recognized, which are represented by the merging and splitting of local topics in different time periods. Results show that the evolution of a major topic usually follows a pattern from adjusting status to mature status, and sometimes with re-adjusting status in between the evolving process. Knowledge transfer happens both within a topic and among topics. Word migration via topic channels has been defined, and three migration types (non-migration, dual-migration, and multi-migration) are distinguished to facilitate better understanding of the topic evolution.

Keywords: Topic evolution; Semantic word shifts; Content analysis

1. Introduction

Topic evolution indicates how a topic is changing over time, including whether it is maturely developed, imports knowledge from other topics, merges or splits into others, as well as which topics are gaining importance or dying out. All these evolutionary characteristics are not only meaningful by themselves, but also informative for researchers for better understanding of the domains. Understanding topic evolution can facilitate the promotion of knowledge transfer within and across domains, and help funding agencies and decision-makers keep track with innovations and knowledge flows. The large volume of publications brings challenges to gaining an overview of a field, but it becomes a great source of ideas with rich context for mining and learning the evolving process of innovations (Ding & Stirling, 2016).

The evolution of topics in scientific domains has been extensively explored in scientometrics and data mining areas. Knowledge Domain Visualization (Börner et al., 2003) has been used to gain an overview of a field, and visualize emerging trends and research frontiers. Topic evolution models (Amoualian et al., 2016) have been developed to discover the emergence of a topic and its content transitions. Although scientists have studied scientific topic evolutions to keep abreast of their fields and related topics, most works mainly investigate the macro level of topic changes such as content transitions, keywords overview and prominent cluster visualizations.

But there exists a strong need for more understanding of topic evolution in areas such as (1) topic trends over time, which indicates the active status and turning points of a research topic, (2) The splitting and merging of topics, and whether the topic has absorbed knowledge from other topics or exported its knowledge to influence others, and (3) the developing status of a major topic, which indicates whether the topic is maturely developed or in the process of dying out. Most existing works focus on macro-level content transition of topics, e.g., the changing of top words of topics in *Science* (Blei & Lafferty, 2006), a general view of the most highly used keywords in *PNAS* publications (Mane & Börner, 2004), and the identification of the most prominent clusters in the mass-extinction research over time (Chen, 2006). The problem of how knowledge transfers (referring to the splitting and merging) and major topic develops (referring to recognizing the developing status of topics in different periods) remain

largely unknown.

Because topics are expressed by collections of words, word changes are closely associated with the evolution of topics. When browsing the evolution of a topic, some words initially appearing as top words disappear in later periods. In practice, a word shifts or extends its semantics while a topic evolves, where the same word can appear in multiple topics with different contexts. We define such shifting context of the same word as "word migration" via topic channels. Specifically, word migration refers to the same word embedded in different topics, which resemble the migration of populations in demographics (Williams & Baláž, 2014), as if the word is an analog of a human population and topics are territories. While tracking the migration of words through topic channels facilitates better understanding of topic evolution, the shifting context of words in association with topics has not been explored to our knowledge.

In this paper, we study topic evolution in a scientific domain through (1) discovering topic trends from documents; (2) identifying evolving dynamics, which presents the splitting and merging of topics, the underlying knowledge transfer among topics, and the developing status of a major topic; and (3) tracing the migration of words via topic channels to facilitate better understanding of topic evolution.

The remainder of this paper is organized as follows. The related work section presents a summary of studies related to topic evolution analysis. The methods section describes the dataset, the method of topic extraction, and our methods for discovering topic trends, evolving dynamics, and migration of words. The results and discussions section presents our understanding of topic evolution in a scientific domain from several perspectives. The conclusion section summarizes the findings and suggests future work.

2. Related work

2.1. Topic trend detection

Studies on topic evolution start with discovering topic trends in temporal documents. During the 1970s, topic trend detection was done by content analysis of published articles (Lounsbury et al., 1979). Topic areas are manually defined by examining article content, where the topic trend is summarized by counting the number of articles in an area in each time period. Content analysis by experts is generally accurate, but with the massive amount of publications available today, topics can no longer be summarized or discovered by human annotation. Today topic detection task can be assisted by clustering or topic modeling algorithms. Topic modeling algorithms (Blei, 2012) are statistical methods that can discover the underlying topic themes that run through massive collections of documents. The intuitive understanding of topic modeling is that a document exhibits multiple topics according to a probabilistic distribution. In topic models, topics are usually generated with a multinomial distribution over words, and each document is represented by a multinomial distribution over these topics.

Most studies to date (Börner et al., 2003; He et al., 2009; Zhou et al., 2006) have generated topic trends over time by counting the number of documents in a topic year by year, which discards the property wherein one document exhibits multiple topics according to a proportion. The exception is the study conducted by Wang and McCallum (2006), which proposed Topics over Time (TOT) to monitor topic trends. In their study, the timestamp of a document is generated by a per-topic Beta distribution, instead of using the publication date

as other works usually do. A topic in this case is represented as a multinomial distribution over words, as well as a Beta distribution over timestamps. The topic trend can thus be revealed through the Beta distribution over timestamps. In this paper, the per-document topic distribution is obtained through topic modeling results to study topic trends over time.

2.2. Topic evolution

In scientometrics, Knowledge Domain Visualization (KDV) is used extensively for Identifying and mapping domains from scientific literature. KDVs are a special kind of information visualization that can depict the structure and evolution of scientific domains, usually created from publications, patents, or grants. According to Kuhn's influential theory on the structure of scientific revolution (1962), the development of science is characterized into phases of normal, crisis, revolution, and the new normal phase. The focus of KDV studies is to detect and monitor such macro level paradigm shifts of science through temporal patterns in scientific networks based on, for example, co-citation, co-author, and co-word relations (Börner et al., 2003).

Most document collections are sequentially organized as temporal streams, and thus characteristics of the corpus such as topic content and topic number are time-evolving. To capture such evolving characteristics, topic models that incorporate timestamps have been developed. One seminal work is the Dynamic Topic Model (DTM) proposed by Blei and Lafferty (2006), where documents are organized into time slices. Documents in each slice are modeled with a K-component topic model, where the detected topics are evolved from the last slice's topics. This model generates a chain-like topic evolution route, which primarily focuses on the content transition of a topic, but does not consider the dynamics of topic correlation. Similar to DTM, most existing topic evolution studies focus on the content transition of individual topics. Since DTM is designed for categorical data, Wang et al. (2012a) extended the model to continuous time, so that it can handle large amounts of time points. Wang et al. (2012b) and Gohr et al. (2009) expanded on the Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) models, respectively, for mining text streams. These models allow new documents to be streaming in, new words to emerge, and old ones to be forgotten. The Dynamic Mixture Models (Wei et al., 2007) is similar to DTM, except it assumes a document dependency based on the mixture topic distribution of the document instead of a Dirichlet prior, and the topic-word dependencies are dropped.

Studies on the content transition of topics do not consider the correlations between topics in terms of splitting and merging dynamics. Topic correlations can be represented by flows. Jo et al. (2011) tried to capture the rich topology of topic evolution inherent in the corpus through citation flows, where topics are identified with their time of appearance via a chronological scan over document contents. Mei & Zhai (2005) studied the evolutionary topic patterns through the use of temporal text mining, where topics in different time periods are extracted by applying PLSA. The topic correlations are measured by the Kullback-Leibler divergence between topics, where the splitting and merging flows between topics can thus be generated according to a threshold. The few studies that consider topic correlations generally view the splitting and merging of topics as a structural change, but seldom discuss the knowledge transfer indicated by the splitting and merging activities or the developing status of major topics.

2.3. Semantic word shifts

Topics are essentially collections of words with semantic functions (Griffiths et al., 2005). Defined as a change of one or more meanings of the word in time (Lehmann, 1993), the shifts of word semantics and its detection has been the focus of much research in recent years. Studies on word semantics over time can be viewed from two perspectives: synonymy detection and polysemy detection. Synonymy detection monitors the use of different words with the same meaning over time (Kenter et al., 2015). Based on a small set of input words in a certain time period, ranked lists of terms for a consecutive series of periods in time would be output. The words in the ranked lists are meant to denote the same concept as the input words. Studied more extensively, polysemy detection monitors different meanings expressed by the same word over time. A word can change semantically in a way wherein new meanings replace the old ones, or acquire additional meaning with the original meaning may still be widely used (Wijaya & Yeniterzi, 2011).

In polysemy detection, distributional semantic models (Gulordava & Baroni, 2011; Hamilton et al., 2016; Kim et al., 2014) are widely used for quantitative measurement. In these models, the similarity between words is measured by vector space models where each word is associated with its context vectors. Existing works have studied the semantic shifts of words mainly to facilitate natural language applications, for example, time-aware query expansion for document retrieval tasks in a historical corpus. In this case, words are studied alone and are not associated with topics.

To get a better sense of the process of topic evolution, the word migration patterns we investigate in this study are closely associated with topics. The semantic shifts of words are represented by the words being embedded within different contexts in different topics.

Migration in all forms is common in the real world. In knowledge-based economy, international immigration of skills and knowledge stimulates the spreading and evolving of technologies (Williams & Baláž, 2014). In textual topics, words serve a similar role as that of human populations, where migration of words across topic boundaries may also indicate communication between topics. Due to the changing context of the migrated words, word migration is not simply a replicate of lexical information, but rather an idea re-creation (Iles et al., 2004). The same word, once it migrates to another topic, is likely to represent new ideas due to being in a different context.

3. Methods

3.1. Data and topic extraction

Information retrieval (IR) is chosen as the target domain. Papers are collected from Web of Science for 1956-2014, making a total of 20,359 documents, with search based on a set of IR-related terms. Search term selection refers to the paper by Xu et al. (2015). The selected document types include article, book, book chapter and proceedings paper. The title and abstract fields are used as the text corpus for extracting topics.

Before extracting topics, all terms are stemmed using the Porter2 stemming algorithm. A stop word list (Yan et al., 2012) is used to filter common words. Words with only one letter or appear less than five times are removed.

The Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) is applied for extracting topics from the corpus. LDA is a three-layer Bayesian model that is now widely used in discovering the latent topic themes in collections of documents. The LDA model represents

each document with a probability distribution over topics, where each topic is represented as a probability distribution over words. For a detailed explanation of the algorithm, refer to, e.g., Blei (2012). The Gensim library (Rehurek & Sojka, 2010) is used for implementing the LDA model, where the parameters are set as the standard value proposed by Gensim.

For this study, we extract five topics in total from the corpus. The selection of topic numbers in LDA models is always an open problem. The widely used metric perplexity usually gives a best number of topics between 80-100. This number is too large for our study, for this study is primarily targeting on the major topics in IR. After checking topics in the top IR conferences, combining with our understanding of IR based on previous studies of our group conducted on authors and communities in the IR field (Ding, 2011; Yan et al., 2012), we consider a topic number between 5-10 is appropriate. Among these values, the choice of five topics gives the best topic coherence. The coherence of topics is evaluated by human judgments (Chang et al., 2009).

We reran our analysis with alternative values of the topic number other than five to test how it affects the evolution structure when changing topic numbers. For the test involves local topics in different time periods, detailed explanations are given in the end of Section 3.3, presented along with the time-window changing test.

3.2. Detecting topic trends

From the topic extraction results, each document is presented with a probabilistic distribution over the five topics. The popularity of a topic over time is calculated by aggregating the per-document topic distribution by year.

Figure 1 presents the example documents from years 2007, 2008, and 2009. Each row in the left table represents one document, and the values are probability distribution over topics that indicate the proportion of the underlying topics a document contains. Documents are grouped by year, and the probabilities of the same topic in the same year are summed up to get the upper-right table. Results are then normalized by dividing the values by the row sum, which is the number of publications in that year. Finally, the topic popularity values are presented in the lower-right table. The sum of each row is 1, and each value is the proportion of the contents in all documents a topic shares in the corresponding year. The higher the proportion, the more popular the topic.

	Year	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Sum year-2007	2007	0.7650	0.0432	0.1866	0.0025	0.0025
	2007	0.1330	0.0013	0.3994	0.0013	0.4646
	2007	0.1428	0.0029	0.7777	0.0734	0.0029
	...	...	...	...	...	...
Sum year-2008	2008	0.0400	0.8393	0.0400	0.0400	0.0403
	2008	0.8143	0.0036	0.1232	0.0036	0.0551
	2008	0.0019	0.7112	0.0019	0.2828	0.0019
	...	...	...	...	...	...
Sum year-2009	2009	0.1576	0.0010	0.8391	0.0010	0.0010
	2009	0.0022	0.0021	0.7371	0.2563	0.0021
	2009	0.0176	0.0013	0.7132	0.2306	0.0371
	...	...	...	...	...	...

Year	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
2007	184.0730	197.0024	165.0664	181.5497	180.3084
2008	229.5481	265.9666	172.4949	182.4320	227.5584
2009	202.3045	209.7801	194.4993	206.8327	222.5834

Year	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
2007	0.2027	0.2169	0.1817	0.1999	0.1985
2008	0.2129	0.2467	0.1600	0.1692	0.2110
2009	0.1952	0.2024	0.1877	0.1996	0.2148

Figure 1. Steps for generating topic popularity

3.3. Discovering evolving dynamics

The evolving dynamics to be investigated include the splitting and merging of the topics, the knowledge transfer between the topics indicated by the splitting and merging, and the

developing status of each topic in different time periods. The entire corpus is divided into six time spans with a five-year interval: 1956-1990, 1991-1995, 1996-2000, 2001-2005, 2006-2010, and 2011-2014. The division of the time span is decided by making the evolving process as detailed as possible based on the premise that each period has sufficient textual content, and the division is approximately in accordance with other research related to the IR domain (Ding, 2011).

From each time span, we extract several topics using the LDA model. For differentiation, we name the time-span topics as local topic, and the five general topics as global topic. The evolving dynamics of the global topics is indicated by the merging and splitting between local topics in adjacent time spans. The detailed methodology is explained as following.

Topic correlations. The correlation between a local topic and a global topic is measured by the cosine similarity between their probability word distribution (Figure 2). The word distribution of each topic is approximately a sparse vector. The cosine similarity performs well on distinguishing the correlations between such sparse vectors, where it highlights the contribution of the top rank words with high probabilities and weaken the noise produced by the words with low probabilities. We first select one of the global topic as the target topic, and then calculate the similarity between each local topic in a certain period and the target global topic. For example, in Figure 2, the target topic is set as global topic 1, and the colored circle indicates the strength of the similarity between the local topic and the global topic. Note that local topic 1 in period 1 and local topic 1 in period 2 do not necessarily represent the same topic. Similarities above 0.5 are further distinguished into three intervals, representing weak (green), medium (yellow) and strong (red) correlations to the global topic. The correlation between local topics in adjacent time spans is also indicated by the similarity between their probability word distributions, where the local topics can form merging and splitting flows. The final similarity intervals are set at (0.5, 0.65], (0.65, 0.75] and (0.75, 1], referring to weak, medium and strong correlations. At first, we decide the intervals by dividing 0.5 by 3 and round to 0.15, which makes the intervals (0.5, 0.65], (0.65, 0.8] and (0.8, 1]. But we found that there are several similarities between 0.77-0.79. We consider such similarity should also belong to the strong level, which makes the final intervals (0.5, 0.65], (0.65, 0.75] and (0.75, 1].

For the number of local topics to be extracted from each time span, we tested several values, and after evaluating each value based on the coherence of terminology and ensuring the presentation of the merging and splitting activity, the final number is fixed to ten.

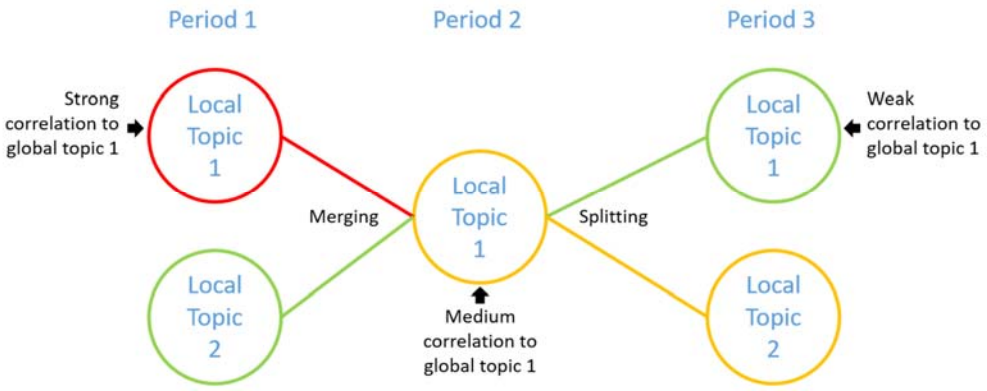


Figure 2. Topic correlation example

Mapping the evolving process. The evolving graph of a global topic mainly presents two kinds of information:

(1) **The developing status of a global topic**, which can be indicated by its correlation with the local topics. A strong correlation shows that the word distribution of the local topic is very similar to the final word distribution of the global topic, which indicates the global topic is in a maturely developed status, and weak and medium correlation indicates a developing or re-adjusting status.

(2) **The merging and splitting flows** that represents the evolving structure and knowledge transfer between topics. The knowledge transfer happened within a global topic and between global topics are reflected by the merging and splitting between local topics in adjacent time spans.

The evolving graph of a global topic is generated through the following steps:

(1) Select a global topic. Find the first local topic which has a strong (red) correlation strength to the global topic. If there is more than one earliest red local topic, keep them all.

(2) Check the similarities between the first red topic/topics and the local topics in the previous time span, and mark out flows between local topics with green, yellow, and red according to the similarity intervals. Then check the time span before the previous time span with the same principle until reaching the first time slice in 1956-1990.

(3) Check the similarities between the first red topic/topics and the local topics in the next time span with the same principle in step 2 until reaching the very last time slice in 2011-2014.

The choice of the global topic number and the time window selection are the two key factors for this study. We reran our analysis with different topic numbers and time windows. In summary, the results are in accordance with the current study, namely, similar conclusions are generated regarding the developing status and knowledge transfer of the IR field.

When the topic number is set to four, two topics (topic 3 and topic 4) of the five-topic results merged into one, the rest remain similar. The merged topic displays mixed themes of topic 3 and topic 4. The local topics with high correlations to the two previous global topics also appears to be highly related to the merged topic. When the topic number is larger than five, the extra topics become less coherent or present general words in the IR field, e.g., *algorithm*, *measure*, *learn*, etc. The evolution process of the extra topics does not have much interactions with other topics, for they have scattered themes. The original five topics still appear in the new extraction when changing topic numbers with similar evolution structures.

Considering the sensitivity of changing the time windows, when the time window gets larger, periods become fewer, and the evolving structure is relatively simplified, where some splitting and merging regarding knowledge transfers between topics are lost, but the developing status of each topic is in accordance with the five-year interval results. In contrast, when the time window gets smaller, more details of the evolving structures are presented.

No matter how the topic number or time window changes, the evolution process of the IR field as a whole exists consistently. Similar conclusions are drawn considering the knowledge transfer activities and the developing status of major topics.

3.4. Tracing the migration of words

A word's migration is examined herein based on its yearly topic distribution. The topic distribution of a word in a document is parameterized by the variational parameter φ in the LDA model (Hoffman et al., 2010). The φ value indicates the likelihood of a word belonging to a topic in terms of a particular document. Each word in a document has five topic φ values corresponding to the five global topics (Table 1). After normalizing by the frequency of the word in the document, the sum of a word's φ values is equal to 1. The same word from two different documents usually has two different sets of φ values. For a selected word, we calculate its average φ value for each topic in each year. The average φ values represent the topic probability distribution for the word in that year. As the φ values change over the years, the word migrates between topics over time.

Table 1. φ values of the word *user* in document *20143*

Document ID	Year	Word	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Sum
20143	2010	user	0.999904	7.22E-06	2.53E-05	6.16E-06	5.72E-05	1

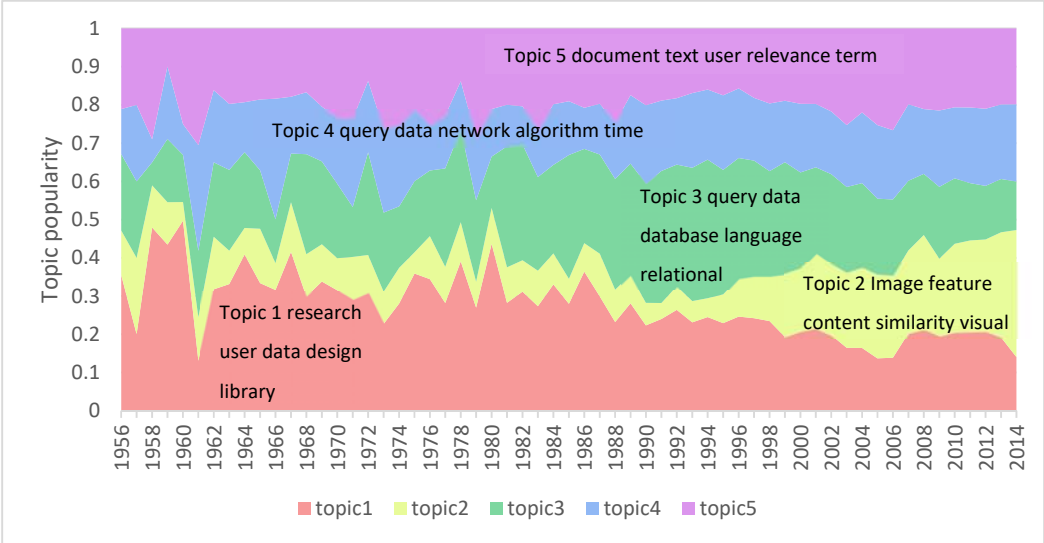
4. Results and Discussions

4.1. Topic trends

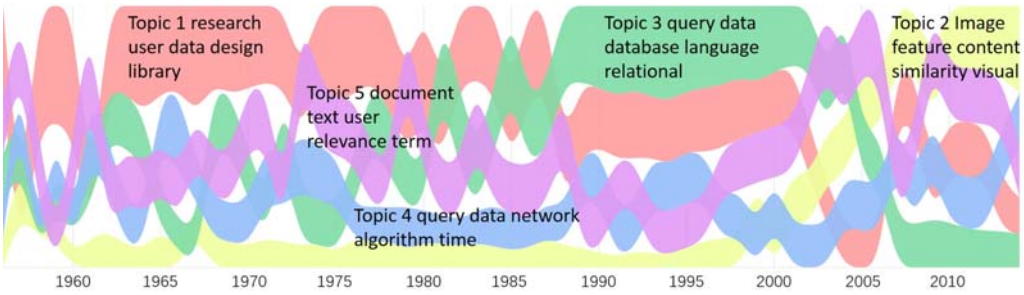
The top ten words with the highest probabilities of the global topics are presented in Table 2. Figure 3 shows the topic trends with two types of displays. Topic popularity is indicated by the proportion it shares in each year. Figure 3(a) uses flow width to reflect topic popularity, where the wider the flow, the more popular the topic. Figure 3(b) uses both width and position of the flow to present topic popularities, where the nearer to the top, the wider the flow, and the more popular the topic.

Table 2. Top ten words in global topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
research	image	query	query	document
user	feature	data	data	text
data	content	database	network	user
design	similarity	language	algorithm	relevance
library	visual	relational	time	term
web	music	semantic	index	query
find	learn	integration	distributed	web
analysis	algorithm	structure	optimize	evaluation
medical	object	object	computing	rank
access	color	knowledge	tree	word



(a)



(b)

Figure 3. Topic trends

All the five global topics fluctuate around the probability value of 0.2, which makes sense because there are five global topics (0.2=1/5). But the changing tendencies are quite different. Topic 1 and Topic 3 have a shrinking or declining tendency in general. Topic 1 (user study) reaches its peak before the 1960s and suddenly reaches a minimum in 1961. But it gains back its popularity soon thereafter, and then gradually loses popularity in the following years, reaching bottom in 2005. Topic 3 (database querying) dominates in the early 1990s, but starts to shrink from around 1994, and sharply goes down all the way to the bottom in the 2010s. Topic 2 (image retrieval) starts to expand after 1995 and reaches the top in the 2010s. Topic 4 query processing) remains approximately stable over time, and slightly expands after 2000. Topic 5 (text retrieval) stays about in the middle before 2000, and gains more popularity around 2005, while topic 1 and topic 3 shrink.

4.2. Evolving dynamics

Evolving graphs of the global topics are presented in Figure 4. The evolving dynamics of each global topic is presented separately, because a certain local topic usually has different correlations to different global topics.

For clarity of strong correlations, we discard some weak flows (green) in the presence of red or yellow flows. The dot-line circle frames in green means the similarity between the local topic and the global topic is below 0.5. The top words of local topics in the evolving graphs are presented in Table 3.

Global topic 1 (Figure 4(a)) focuses on user-oriented problems, covering online information-seeking behavior, use of digital resources such as digital library by research scholars, and user information needs, especially for health information search.

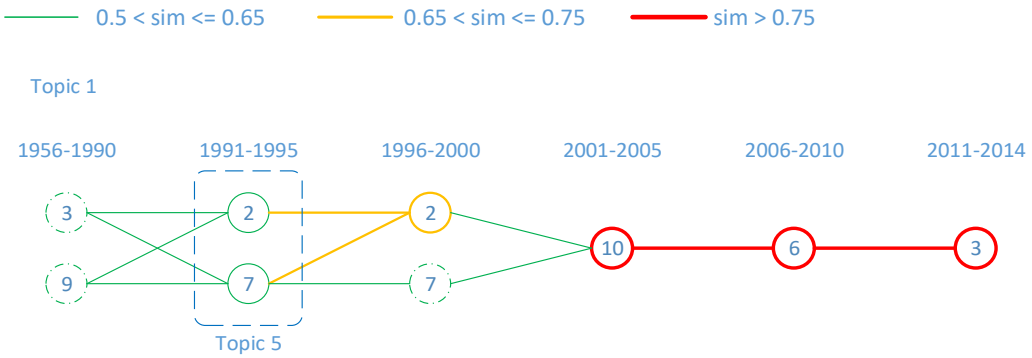


Figure 4(a). Global topic 1

Local topic 10 in 2001-2005 is the first local topic that has a strong correlation to global topic 1, indicating the topic gets maturely developed in this period. Before 2001-2005, the topic is in a developing or adjusting state, referring to the correlations between the local topics and global topic 1 are no stronger than the medium level. The main theme of global topic 1 is steadily passed along throughout 2006-2010 and 2011-2014, which can be confirmed from both the strong similarity connections (red).

Its evolving process based on the evolving structure and document contents is interpreted as follows: In 1996-2000, when the arrival of internet results in the development of information retrieval, use of online resources and electronic information systems begins to form a stand-alone research area. In this period, the local topics start to pay attention to user search goals, and try to conduct systematic evaluation of web-based search engine performance based on real user behavior and information needs. The change of libraries from the traditional print to digital forms are noted. During this period, user concern about medical information retrieval is studied as a notable sub-topic. For example, numerous studies are conducted on accurate health information retrieval and intelligent medical information filtering for specific health information needs. In the next period of 2001-2005, which is when global topic 1 gets maturely developed, the studies focus more on user interactions and information needs in both web search and digital library environments. Medical information retrieval starts to involve patient care and gradually proceeds to clinical decision support.

Note that local topics 2 and 7 (both related to document and text retrieval, top words presented in Table 3) in period 1991-1995 have a much stronger similarity to global topic 5 text retrieval (similarity 0.69 and 0.77, medium and strong) than to global topic 1 user study (0.55 and 0.52, both weak). It can thus be inferred that the generation of global topic 1 is originally influenced by the knowledge of global topic 5. This can be confirmed by examining the top words of local topics relating to global topic 1 before 1996 in Table 3, and the top words of global topic 5 in Table 2. Table 2 indicates that global topic 5 has a strong relation to the word *document*, so it is not surprising that *document* plays the most important role in global topic 1's local topics before 1996.

Table 3. Top words of local topics

Time span	Local topic	Top words

1956-1990	3	document - analysis - cluster - language - translation
	6	query - language - database - relational - distributed
	7	query - distributed - thesaurus - computer - control
	9	document - data - query - index - theory
	1	database - data - object - query - manage
	2	document - text - network - database - research
	6	data - query - user - compute - database
	7	document - text - user - index - knowledge
	8	query - language - database - relational - data
	1	query - database - data - object - language
	2	user - document - web - relevance - research
	5	query - database - data - time - algorithm
	7	knowledge - learn - case - user - network
	8	image - feature - content - similarity - database
	9	query - language - relational - logic - express
	10	document - index - text - data - structure
	1	query - language - data - xml - database
	2	image - feature - content - color - database
	5	text - document - answer - question - word
	7	document - relevance - user - query - learn
	8	query - data - video - time - index
	10	research - knowledge - data - user - analysis
	1	data - query - database - xml - structure
	2	document - web - text - user - semantic
	3	query - algorithm - object - graph - spatial
	4	image - feature - content - visual - color
	6	research - user - design - find - evaluation
	9	query - language - term - ontology - index
	10	query - data - network - sensor - index
	2	semantic - ontology - word - concept - annotation
	3	research - web - user - library - article
	6	query - data - network - algorithm - index
	7	data - database - query - language - web
	8	document - term - query - text - topic
	10	image - feature - similarity - visual - content

Global topic 2 (Figure 4(b)) has been a unique topic among the five global topics. This topic forms in 1996-2000 without importing flows from previous periods, and has been a standalone topic from the very beginning. Its main theme centers on multimedia information retrieval, especially image retrieval. It does not interweave much with other topics, because its research theme is coherent and unique in the domain. From its generating period to the most recent period, it proceeds in an exclusively stable way, focusing on its main research theme.

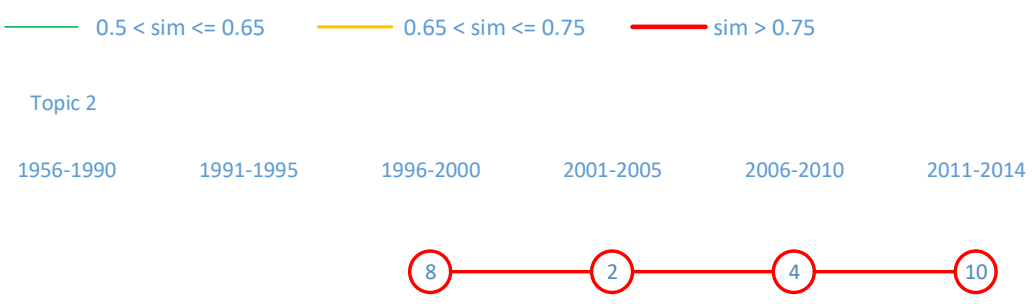


Figure 4(b). Global topic 2

Global topics 3 and 4 are the two topics that interweave with each other the most. These two topics both study queries for structured data sets. Global topic 3 mainly deals with traditional query processing for relational and object-oriented databases. Global topic 4 primarily focuses on distributed query processing for spatial networks and communication networks.

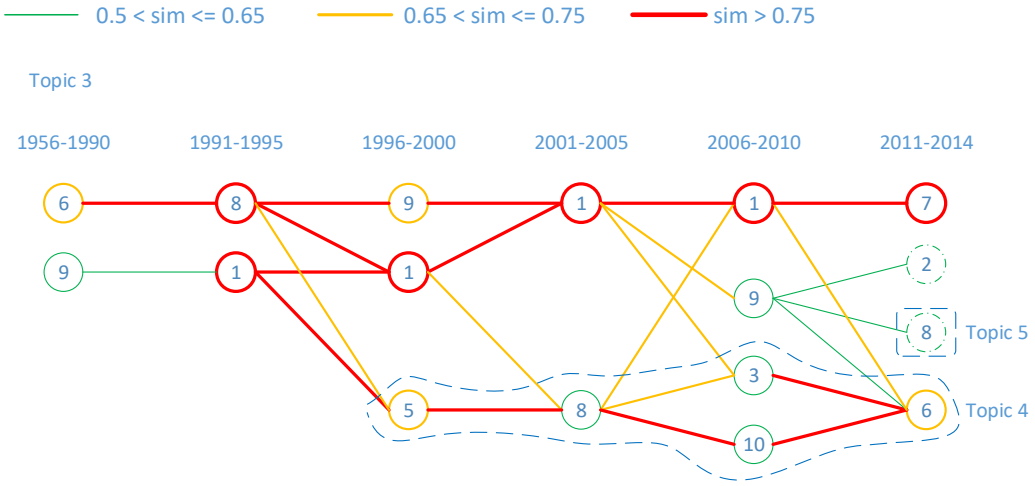


Figure 4(c). Global topic 3

The main theme of global topic 3 forms quite early (Figure 4(c)). In 1956-1990, local topic 6 is the only local topic that has a medium similarity strength in relation to any of the global topics. In other words, when the other global topics are still in an adjusting status in the first period, global topic 3 already stands out as a half-way matured topic.

Local topic 6 and 9 in 1956-1990 flows into local topic 8 and 1 in 1991-1995 respectively. And in 1991-1995, local topic 1 and 8 both split into several following topics. From the splitting flows between 1991-1995 and 1996-2000, we can see that global topic 4 is originally derived from global topic 3, for the local topics in the lower part of the graph are all in strong correlation with global topic 4 (see Figure 4(d)), but only weak or medium correlation with global topic 3.

The evolving process of global topic 3 regarding knowledge transfers based on the splitting and merging structure is interpreted as follows: The topic gets maturely developed in 1991-1995, and in this period, the knowledge of local topic 8 and 5 involving query optimization in distributed database systems merge into local topic 5 (real-time and distributed query-processing systems) in 1996-2000. Other parts discussing architectural features and algebra for object-oriented and relational database continue to develop through the main flows of global topic 3. In 2001-2005, local topic 1 splits into three parts: one keeps flowing within global topic 3, another part discussing tree transducers and transformations for xml queries flows into global topic 4, and a third part relating to ontology-based searching flows into local topic 9 (cross-language query suggestion) in 2006-2010, which finally merges into global topic 5, represented by local topic 8 (with the similarity to global topic 5 around 0.84, strong) in 2011-2014.

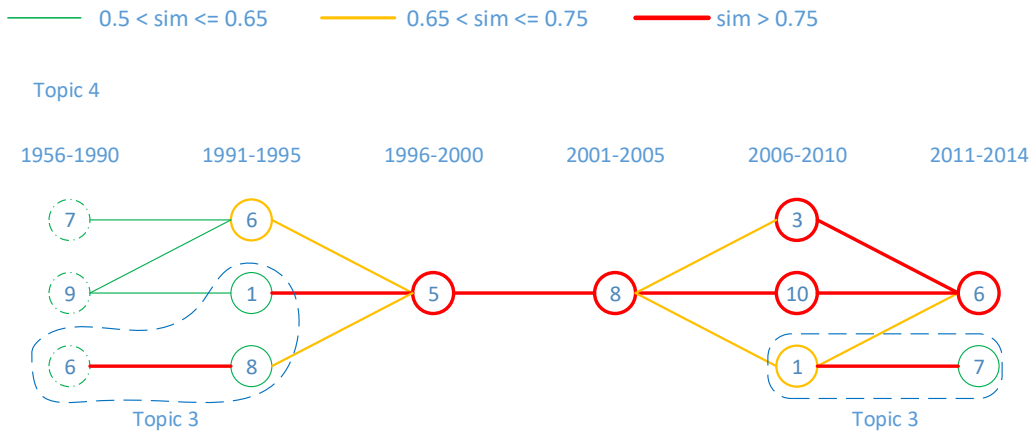


Figure 4(d). Global topic 4

Global topic 4 (Figure 4(d)) becomes mature in 1996-2000, referring to local topic 5 in 1996-2000. Local topic 5 in 1996-2000 absorbs part of the knowledge from global topic 3 (referring to local topics 1 and 8 in 1991-1995) and local topic 6 in 1991-1995. During this period, global topic 4 develops its own theme on execution strategies for real-time database systems with timing constraints and distributed query-processing systems, such as parallel database systems. The research focuses of local topic 5 in 1996-2000 steadily passes to local topic 8 in 2001-2005, and then splits into local topics 1, 3, and 10 in 2006-2010. These three local topics re-merge into local topic 6 in 2011-2014, which mainly studies data storage, indexing, and complex queries in distributed environments, such as peer-to-peer systems and sensor networks. In general, global topic 4 at first inherits knowledge regarding query optimization in distributed database systems from global topic 3 through local topic 5 in 1996-2000, and then returns parts of its knowledge regarding spatio-temporal databases back to global topic 3 through local topic 1 in 2006-2010.

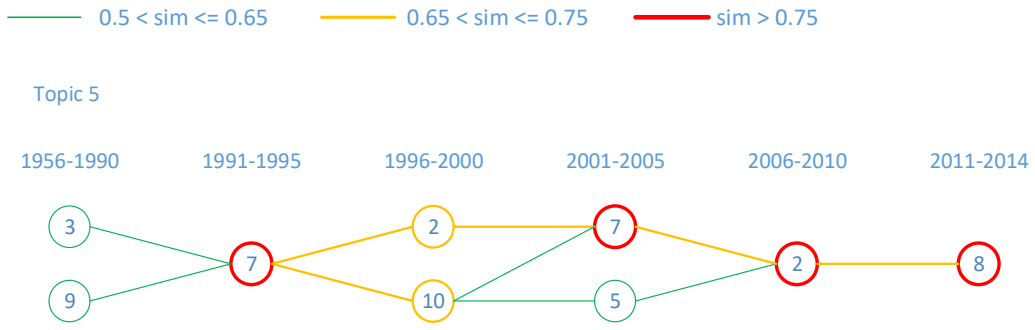


Figure 4(e). Global topic 5

Global topic 5 (Figure 4(e)) studies text retrieval for unstructured documents, which involves document indexing and terminology processing problems, such as term disambiguation, query expansion, and cross-language retrieval. A significant evolving characteristic of global topic 5, which is quite different from the other global topics, is that there haven't been strong connections between local topics in adjacent time spans. In other words, although the similarity strength between the local topics and global topic 5 is fairly strong (though it fluctuated in 1996-2000), the similarity strength between local topics in

neighboring time spans are no stronger than the medium level. This reflects that global topic 5 is a rather active topic, for there is always new knowledge introduced or created in this area.

In 1991-1995, local topic 7 discusses the document indexing problem in general. Studies in local topic 7 cover document representation, text categorization and interpretation, relevance judgment, and evaluation of retrieval performance. These studies have been a technical foundation in 1996-2000 for local topic 2, conducting user-centered experiments, which becomes the starting point of global topic 1 (user study). Local topic 2 in 1996-2000 also discusses evaluation of retrieval performance, which is consistent with the studies in 1991-1995, while local topic 10 leads the research towards term-level consideration, introducing ideas such as latent semantic indexing for documents, word segmentation, and phrase mining methods. Still in 1996-2000, hypertext retrieval also draws wide attention, where scholars start to consider links between documents, not only the textual information in the documents.

In summary, period 1996-2000 starts to view retrieval problems from the term-level perspective, but also displays several other concerns. For its blended research themes, this period can be treated as a re-adjustment phase for global topic 5. In 2001-2005, studies mainly focus on improving relevance feedback. The similarity to global topic 5 in 2001-2005 returns to the strong level (red), which indicates global topic 5 has developed a new mature status with new knowledge imported into it. In the following periods, the link strength between local topics still does not surpass the medium level (yellow), yet the similarity to the global topic remains strong, which indicates each period has some new knowledge created or imported, but also centers on the main theme of global topic 5. In 2006-2010, personalized web search is promoted. Applications include user modeling, personalized document clustering, and personalized detection of fresh content. Semantic and contextual search gain more importance, combining with text retrieval. New methods are developed for automatic extraction of key phrases, hierarchical document clustering, and link structure analysis. In 2011-2014, multilingual retrieval is developed, and semantic search goes further into sentence level to include local context.

4.3. Migration of words

Topics are essentially collections of words, and tracking the migration of words through topic channels facilitates better understanding of topic evolution. The migration of words is examined herein after 1990, because all global topics are formed no earlier than the time span of 1991-1995, and the field of information retrieval starts to thrive in the 1990s. The topics in this section are all global topics, which is called “topics” for simplicity. Primarily, the top ten words with the highest probabilities in each topic are examined. The top words are all high-frequency words, which ensures sufficient co-occurrence data to get reliable migration results, and avoiding the random appearing of low-frequency words caused by the long tail frequency distribution.

Considering the number of topics between which a word migrates, there are three types of migration patterns in general: non-migration, dual-migration, and multi-migration.

4.3.1. Non-migration

Non-migration words refer to those that always belong to only one topic over time. Typical examples are *image* in topic 2, referring to image retrieval; and *document* in topic 5, referring to document indexing (Figure 5). Non-migration words are strongly bounded to a

particular topic, which usually represents the core research theme of the assigned topic.

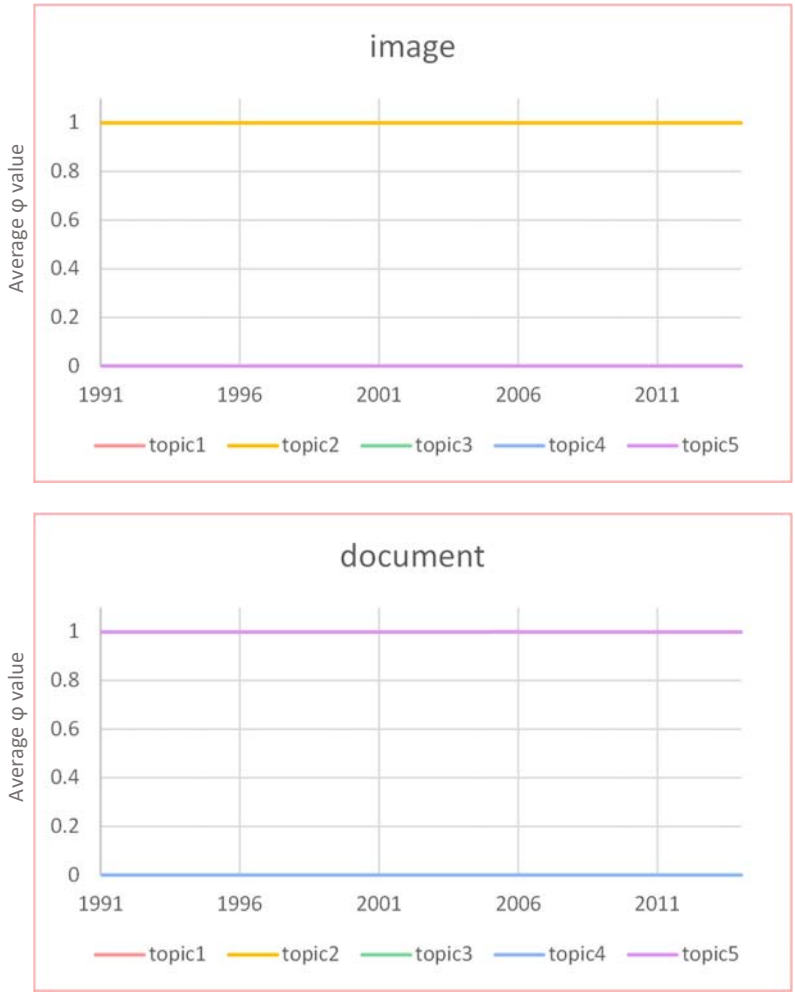


Figure 5. Non-migration words with consistent probability

4.3.2. Dual-migration

Dual-migration words are those that are obviously distinguishable solely in two topics. Curves of the two topics are usually symmetric, as one goes down and the other goes up. Words presented in Figure 6 belong to the dual-migration type.

The word *similarity* migrates between topics 2 and 5. The main theme of topic 2 is multimedia retrieval, notably for image retrieval. Topic 2 is not formed until 1995, so before 1995, *similarity* is mostly assigned to topic 5 in the context of content similarities between texts and documents, as the main theme of topic 5 is text retrieval. The exception is 1993, when *similarity* is assigned to topic 2 with a higher average probability of 0.6, and a lower average probability of 0.2 for topic 5. This is because in 1993, several studies emerge on image database systems for supporting the storage and retrieval of images by content, which results in the word *similarity* being assigned to topic 2 with a higher average probability. As topic 2 begins to form, *similarity* shifts its context to similarity measure for images or other multimedia sources, and is then assigned to topic 2 with a higher probability thereafter. The word's average probability in topic 5 does not die out, however, and maintains a value of around 0.2. The meaning of the word *similarity* extends throughout the migration, and finally

becomes restricted mainly within two topics.

The word *language* migrates between topics 3 and 5. When in topic 3, it is related with query language in databases. When in topic 5, it is usually placed in the context of cross-language retrieval. Since the dual relationship is always compensating in order to add up to the sum of 1, when the probability for topic 3 is declining, topic 5 is increasing. These tendencies indicate the migration of *language* from topic 3 to topic 5, for it is studied more in cross-language problems rather than database query-language contexts in recent years.

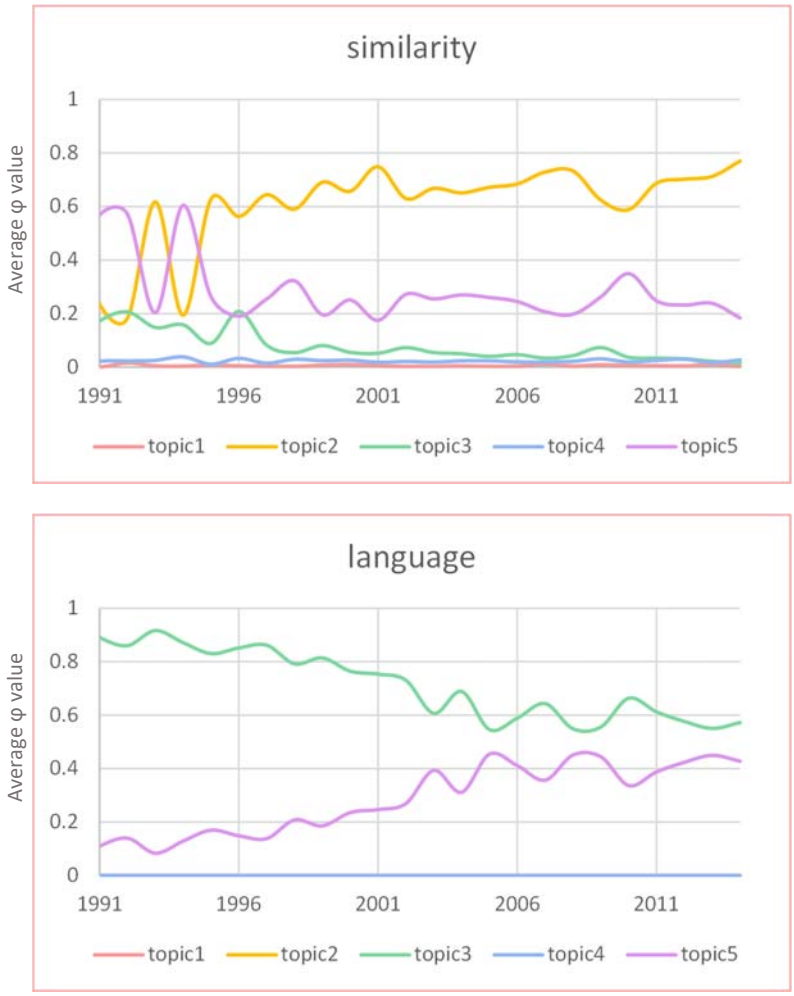


Figure 6. Dual-migration words

4.3.3. Multi-migration

Multi-migration refers to a word that is assigned to more than two topics during topic evolution. There are usually multiple underlying ideas represented by these words, so they display migration among multiple topics, e.g., *user* (Figure 7) starts with an assignment to three topics and ends up in being studied more in topic 5 on personalized web search and also in topic 1 on end-user behavior.

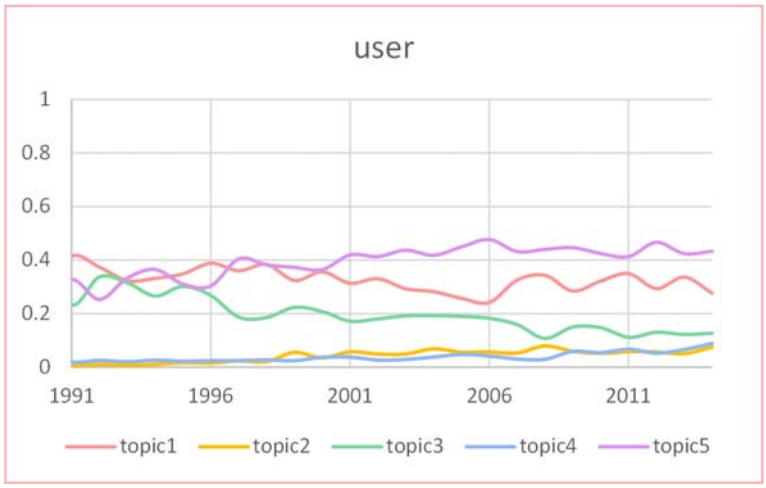
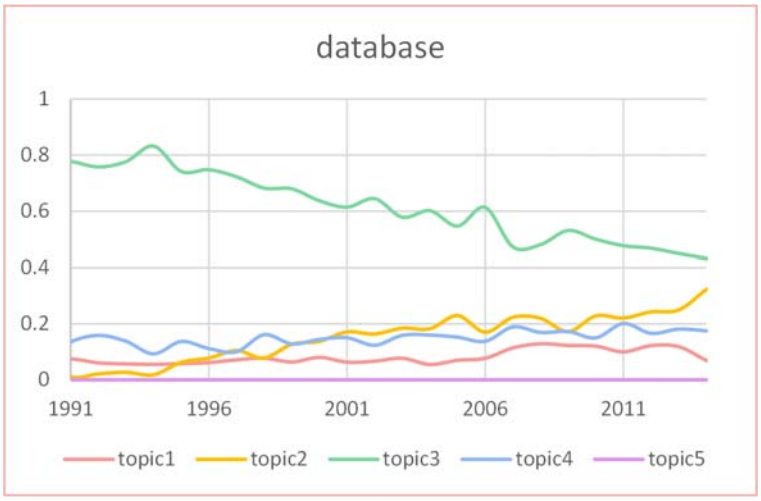


Figure 7. Multi-migration word *user*

The probability of topic assignment represents the likelihood of a word being studied within a certain topic. Multiple-word display of a declining tendency in one particular topic can indicate that the topic is dying down in popularity. This connection is especially obvious in multi-migration scenarios, where a word loses its popularity in one context due to the shrinking of the topic and permeates into several other topics. A typical example is in topic 3, where Figures 8 shows three multi-migration words with a declining probability.



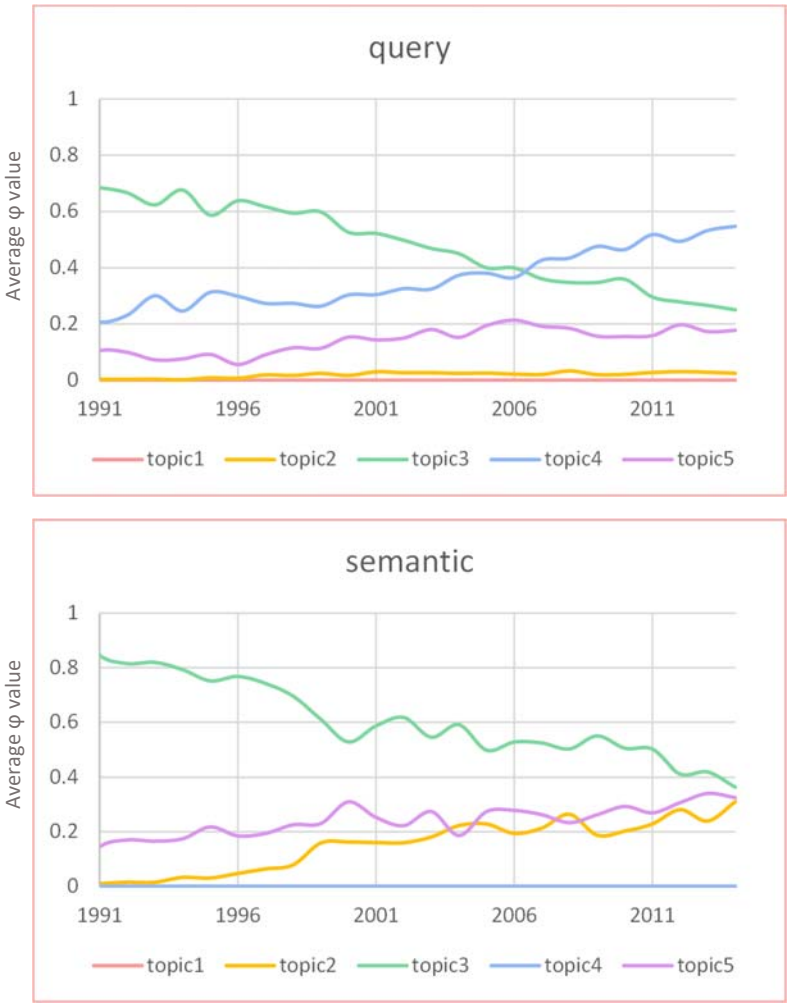


Figure 8. The migration of words from Topic 3 to other topics

Database migrates from topic 3 to topics 2, 4 and 1, especially to topic 2, image and multimedia retrieval, for its curve continues going up in the 2010s. The word's popularity over all documents in each year is also declining (Figure 9). A word's popularity represents the proportion of papers that mention the word in the title field. If there are ten papers in a certain year, and three of them have *database* in their titles, then database's popularity in that year is 0.3. A declining popularity suggests that the research on database querying itself has been well developed, and no longer draws much attention in general. Database techniques becomes a fundamental basis for information retrieval tasks, and merge into various contexts, combining with specific applications such as image retrieval.

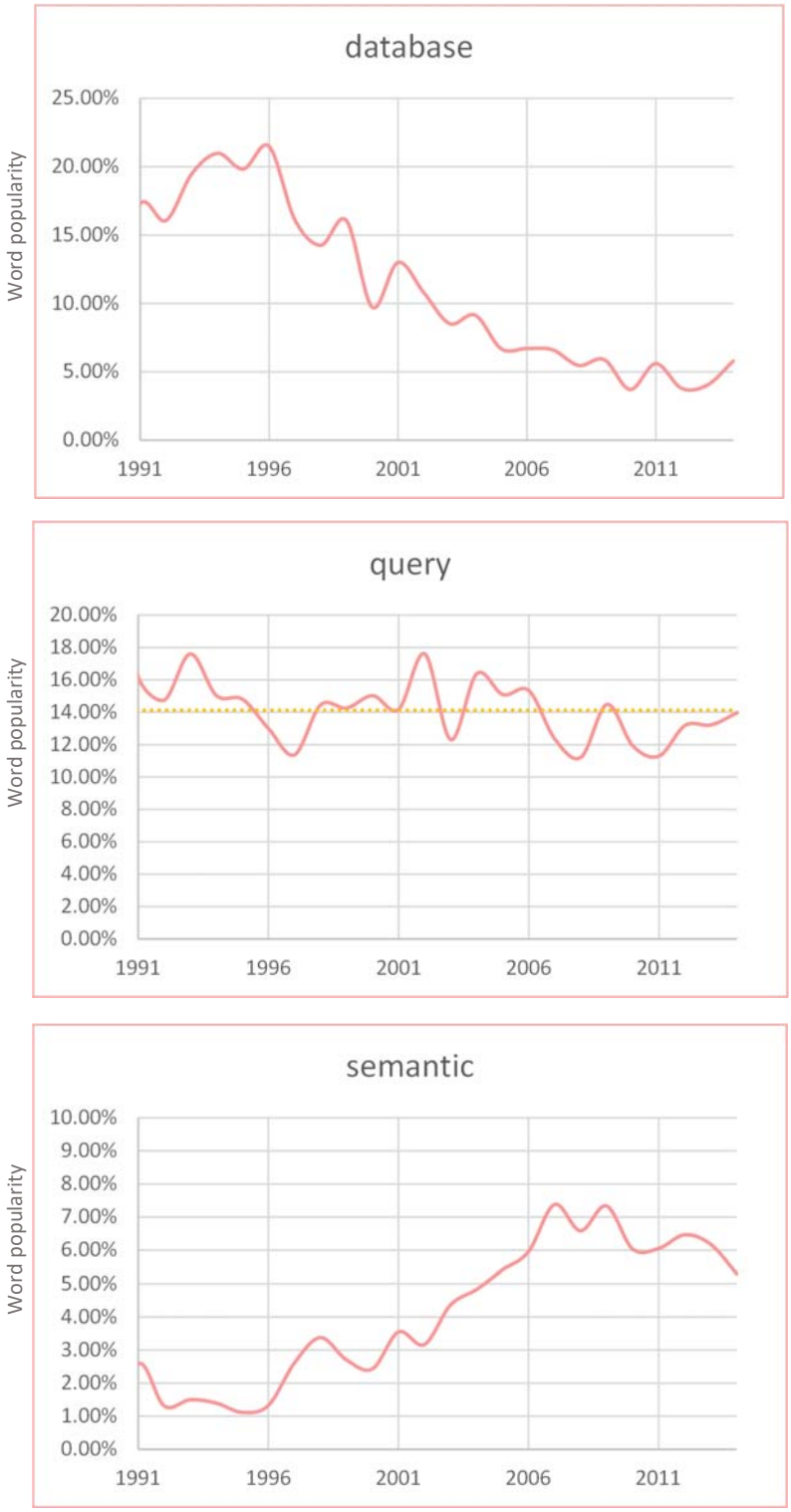


Figure 9. The diachronic popularity of words in Topic 3

The word *query* mainly migrates from topics 3 to topic 4, while its probability in topic 5 remains around 0.2 over time. Its popularity does not change much (Figure 9). On average, 14 percent of paper title fields contain *query*. The stability in its popularity shows that the word is still very important in the field of information retrieval, but in the local context, its focus

migrates from traditional query processing (topic 3) to advanced distributed query processing (topic 4).

The popularity of *semantic* (Figure 9) increases from one percent in the 1990s to around six to seven percent in the 2000s, suggesting an increasing attention paying to semantic studies in the IR field. For the migration status (Figure 8), it is at first assigned mostly to topic 3 under the context of semantic query language, and then gradually migrates to topic 5 with the context of semantic text retrieval, and topic 2 with the context of image semantic retrieval.

The dying down of a topic's popularity and the words migrating to other topics occur simultaneously, e.g., as topic 3 shrinks, its top words migrate to other topics to develop new semantics. Some words like *database* may lose its popularity in general, for the innovation originally attached to the word is less interesting over time. Other words may retain a stable proportion or increase in word popularity, such as when the migration changes its semantics and the word becomes used in new contexts.

5. Conclusion

This study analyzes topic evolution patterns in a scientific domain through investigating the topic trends, evolving dynamics, and migration of words during topic evolution. Findings are listed as follows.

In topic trend detection, the shrinking and expanding tendencies of topics are examined. While other studies commonly use the yearly document numbers to reflect topic trends, our analysis based on the per-document topic distribution reserves the property wherein one document exhibits multiple topics according to a proportion, and gives a clear picture of the temporal popularity and the turning points of the topics.

The evolving dynamics of the phenomenon are investigated from the perspectives of the splitting and merging of topics, knowledge transfer between topics, and the developing status of major topics. Two types of topics are extracted, consisting of five global topics and ten local topics in each of the six time spans. The correlation between topics is indicated by the similarity strength. The similarity strength between the local topics and a global topic reveals whether the global topic has been maturely developed. The splitting and merging of local topics indicates the existence of knowledge transfer within a global topic or between global topics. The evolution of a global topic usually follows a pattern of starting with an adjusting status and gradually standing alone. In the adjustment-period status, a global topic may absorb knowledge from other global topics to generate its own themes. After it becomes mature, the global topic may export its knowledge to impact other global topics. There are exceptions to this, such as global topic 2 (image retrieval), which does not interweave much with other topics. As the research themes in global topic 2 are coherent and unique in the domain, the topic does not have much similarity or knowledge transfer to other topics. The developing status of a global topic can go from mature to a re-adjusting, and then further develop into a new mature status, such as global topic 5 (text retrieval).

Word migration is examined herein based on the average per-word topic distribution over time. The migration patterns can be summarized as non-migration, dual-migration, and multi-migration. A non-migration word is strongly bounded to a particular topic, and usually represents the core research theme of the topic to which it belongs. A dual-migration word migrates between two topics. The probabilities of the word in the two topics are usually

symmetric, for if a word is studied more in the context in one topic, it will be discussed less in another. A multi-migration word expresses multiple semantics that can be studied in multiple topics. When its probability in one topic goes down, it migrates to several other topics to develop new semantics. When several words all exhibit a declining probability tendency in the same topic, it indicates that the topic is dying down in popularity, where this connection between the words and the topic is more obvious in the multi-migration scenario.

The limitation of this study is the discrete time span, which is pre-decided. But no matter how the time span is determined, the knowledge transfer between global topics and the developing status of the global topics should be consistent, because the facts within its evolution usually exist objectively.

There are several interesting directions this research could take. The research can be improved by using a larger data set and analyzing a larger discipline such as the field of Artificial Intelligence. A larger discipline will present more complicated evolving structures regarding knowledge transfers between topics, and the migration activities of words should be more obvious, when the difference between topics becomes more significant. The migration of words can be further explored by analyzing the specific contexts in which they are embedded during the topic evolution, such as the changing group of words that co-occur most with the word in different periods. Furthermore, other knowledge entities such as authors can also be studied in association with topic evolution. The research interest shifts of an author or the research theme shifts of a scientific community can also be expected to facilitate enhanced understanding of scientific topic evolution.

Acknowledgements

This work is funded by the National Natural Science Foundation of China (Grant No. 71420107026 and No. 71704138).

References

Amoualian, H., Clausel, M., Gaussier, E., & Amini, M.-R. (2016). Streaming-LDA: A Copula-based Approach to Modeling Topic Dependencies in Document Streams. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 695–704). New York, NY, USA: ACM.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120). ACM.

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

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). Curran Associates, Inc.

Chen, C. (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.

Ding, Y. (2011). Topic-based PageRank on author cocitation networks. Journal of the American Society for Information Science and Technology, 62(3), 449–466.

Ding, Y., & Stirling, K. (2016). Data-driven Discovery: A New Era of Exploiting the Literature and Data. Journal of Data and Information Science, 1(4), 1–9.

Gohr, A., Hinneburg, A., Schult, R., & Spiliopoulou, M. (2009). Topic Evolution in a Stream of Documents. In *SDM* (Vol. 9, pp. 859–872). SIAM.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17* (pp. 537–544). MIT Press.

Gulordava, K., & Baroni, M. (2011). A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (pp. 67–71). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv:1605.09096 [Cs]*.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 957–966). ACM.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Iles, P., Ramgutty-Wong, A., & Yolles, M. (2004). HRM and knowledge migration across cultures: Issues, limitations, and Mauritian specificities. *Employee Relations*, *26*(6), 643–662.

Jo, Y., Hopcroft, J. E., & Lagoze, C. (2011). The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web* (pp. 257–266). ACM.

Kenter, T., Wevers, M., Huijnen, P., & de Rijke, M. (2015). Ad Hoc Monitoring of Vocabulary Shifts over Time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1191–1200). New York, NY, USA: ACM.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *arXiv:1405.3515 [Cs]*.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lehmann, W. P. (1993). *Historical Linguistics: An Introduction* (3 edition). London ; New York: Routledge.

Lounsbury, J. W., Roisum, K. G., Pokorny, L., Sills, A., & Meissen, G. J. (1979). An analysis of topic areas and topic trends in theCommunity Mental Health Journal from 1965 through 1977. *Community Mental Health Journal*, *15*(4), 267–276.

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5287–5290.

Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198–207). ACM.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks (pp. 45–50).

Wang, C., Blei, D., & Heckerman, D. (2012a). Continuous time dynamic topic models. *arXiv Preprint arXiv:1206.3298*.

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). ACM.

Wang, Y., Agichtein, E., & Benzi, M. (2012b). TM-LDA: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 123–131). ACM.

Wei, X., Sun, J., & Wang, X. (2007). Dynamic Mixture Models for Multiple Time-Series. In *Ijcai* (Vol. 7, pp. 2909–2914).

Wijaya, D. T., & Yeniterzi, R. (2011). Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web* (pp. 35–40). New York, NY, USA: ACM.

Williams, A., & Baláž, V. (2014). *International Migration and Knowledge*. Routledge.

Xu, J., Ding, Y., & Malic, V. (2015). Author Credit for Transdisciplinary Collaboration. *PLOS ONE*, *10*(9), e0137968.

Yan, E., Ding, Y., Milojević, 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.

Zhou, D., Ji, X., Zha, H., & Giles, C. L. (2006). Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 248–257). ACM.