Scientometrics (2023) 128:1567–1582
https://doi.org/10.1007/s11192-023-04642-4

An embedding approach for analyzing the evolution of research topics with a case study on computer science subdomains

Seyyed Reza Taher Harikandeh[1] · **Sadegh Aliakbary**[1] · **Soroush Taheri**[1]

Received: 14 February 2022 / Accepted: 19 January 2023 / Published online: 31 January 2023

Abstract
The study of topic evolution aims to analyze the behavior of different research fields by utilizing various features such as the relationships between articles. In recent years, many published papers consider more than one field of study which has led to a significant increase in the number of inter-field and interdisciplinary articles. Therefore, we can analyze the similarity/dissimilarity and convergence/divergence of research fields based on topic analysis of the published papers. Our research intends to create a methodology for studying the evolution of the research fields. In this paper, we propose an embedding approach for modeling each research topics as a multidimensional vector. Using this model, we measure the topic's distances over the years and investigate how topics evolve over time. The proposed similarity metric showed many advantages over other alternatives (such as Jaccard similarity) and it resulted in better stability and accuracy. As a case study, we applied the proposed method to subsets of computer science for experimental purposes, and the results were quite comprehensible and coherent.

Keywords Topic evolution · Topic embedding · Scientometrics · Informetrics · Data mining · Similarity metrics

Introduction

Nowadays, numerous scientific papers are published daily and the volume of them has increased over last years (Fernández-Isabel et al., 2020). Multiple research areas exist in the scientific environment, and every researcher works on one or more research topics. Moreover, these research topics are not independent of each other. Recently, scientific

✉ Sadegh Aliakbary
s_aliakbary@sbu.ac.ir

✉ Soroush Taheri
so.taheri@mail.sbu.ac.ir

Seyyed Reza Taher Harikandeh
s.taherharikandeh@mail.sbu.ac.ir

[1] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

collaborations were found to play an essential role in the scientific community, and investigating these collaborations attracted much attention (Xia & Liu, 2015). Some research works, particularly interdisciplinary studies, involve diverse research disciplines. As a result, the correlation between research topics is not considered static and may evolve over time. For example, machine learning and data mining are research areas considered as subsets of "Computer Science" (CS). However, nowadays, not only do they contribute to other CS research topics, but they also have found many applications in non-CS research. Therefore, we can study the situation of each research topic among all research areas, and we can analyze trends, changes, and evolution of research topics.

In this study, a "research topic" refers to the "Field of Study" (FoS), defined by the Microsoft Academic Graph. The FoS is a classification of publication records created using entity filtering and iterative graph link analysis (Jung & Yoon, 2020). However, research areas change over time and are neither constant nor static in their graph network. Therefore, the study of "topic evolution" concentrates on tracking topic developments and changes over the years. These changes may include the emergence of new topics, moving topics away from each other, getting closer, or the death of a topic (Chen et al., 2017). Topic evolution studies can help decision-makers such as politicians and administrators to understand the trends of research changes better and make more solid decisions (Qian et al., 2020). Moreover, it also helps researchers to enhance their scientific path and gain better insights that are unknown about their desired research directions (Evans & Rzhetsky, 2011). One of the disciplines that assists the studies on this issue is Scientometrics, also known as the "science of science", which studies the quantitative aspects of science (Van Raan, 1997). The topic evolution analysis is one of the Scientometrics sub-fields that studies the evolution of scientific topics over time. Due to the rapid growth of the sources of information in today's world, the study of topic evolution has received significant attention in recent years (Song et al., 2014).

Due to the importance of computer science in recent research and its interdisciplinary nature, this paper aims to propose a methodology for studying topics evolution in computer science. The results and outcome of such a study have many applications, including career selection of researches, development of interdisciplinary researches, funding, and assessment of research projects (Taheri & Aliakbary, 2022). This study employed an embedding approach to model each Field of Study (FoS) as a vector of real numbers representing a group of researchers' shared research interests. In this context, "embedding" means to map a concept into a finite numeric representation. Nowadays, embedding methods have found many applications in machine learning tasks such as natural language processing (NLP) and graph processing (Masood & Abbasi, 2021; Guo & Caliskan, 2021). Various Fields of Study (FoSs) are transformed to the vector space by employing our proposed embedding method. Consequently, each FoS can be represented in a fixed-sized low-dimensional numeric vector. One of the advantages of mapping FoSs into the vector space is to ease mathematical operations and calculate the similarity measures on different FoSs.

This paper proposes a method for tracking research topic changes over time based on the published scientific articles. Unlike most of the existing methods, we will not directly utilize the text of the papers. Instead, we take advantage of each article's relative FoSs that are already provided in the dataset. Therefore, the similarity of different FoSs is computed based on the proposed embedding of the research topics, which is much less expensive than text-based natural language processing methods. We can investigate the trend of changes in different sub-fields of computer science research in recent years based on our proposed topic embedding method. It will be seen that our proposed method has the ability to calculate the distances between topics that have no relations to each other. Instead

of equating the distance for these topics (with no relation), we use a method in which the distance between each topic pair is significantly determined.

The remaining sections of this paper are structured as follows: in "Related works" section, we study the literature and overview the related works. Then in "Methodology" section , the proposed method is presented. in "Results" Section reports the results of this research, and finally, the paper is concluded in "Conclusion"section , where the future works are also suggested.

Related works

Many approaches have been adopted for the study of topic evolution patterns. This section briefly surveys existing methods related to topic evolution analysis, topic modeling, and topic embedding.

Topic evolution analysis methods seek to determine how research topics change over time. Jung and Yoon (2020) used the change in similar authors' interests to measure the evolution in research topics. In another work, Qian et al. (2020) defined a hierarchical topic model as a tree in which each node represents a research topic, and its children are subtopics of that node. Therefore, the changes in children's structure over time denotes the evolution of the parent topic. The networks based on articles or authors can also be utilized for topic evolution analysis. For instance, Krenn and Zeilinger (2020) presented an algorithm based on semantic networks that are built from the published scientific literature. In this network, vertices denote concepts (topics), and edges indicate the relationship between concepts. This research applies link prediction methods to investigate what concepts will be related to each other in the future. Many related methods are also based on the analysis of the article citations. For example, Kay et al. (2014) used patent co-citation frequencies to find important and trending technologies, and He et al. (2009) proposed an iterative topic evolution learning framework using citation relationships between articles to examine the evolution of topics. Articles keywords are one of the features that can be used in this area. To investigate the formation of interdisciplinary areas, Jian et al. analyzed the evolution of the keywords (Xu et al., 2018). Detecting emerging topic in the future is one of topic evolution goals. Liang et al. (2021) proposed a method for predicting emerging topics. They defined a popularity score for topics under study to forecast emerging topics in the future.

It is worth noting that topic evolution analysis methods are not restricted to analyzing scientific research topics. They have found several applications in social media mining and studying opinion formation patterns. The rise of social media and content-publishing platforms created an excellent opportunity for the study of topic evolution in order to track subjects, or topic shifts on social media (Sayyadi et al., 2009; Becker et al., 2009). For instance, Huang et al. examined the changes of educational topics in social media (Huang et al., 2020), and Alam et al. studied social topic changes based on hashtag co-occurrence (Alam et al., 2017). Furthermore, Kalyanam et al. (2015) performed a new approach based on topic discovery and evolution (TDE) in social media contexts. They adopted a model based on non-negative matrix factorization (NMF) that merges social context and textual information for news articles on Twitter. Allan (2002) presented one of the pioneering research works in topic evolution for tracking new events change in the stream of news stories. This method is basic research in the topic evolution area, which focuses on studying topics evolution based on text streams. Emerging phenomena in today's world have also provided an opportunity for topic evolution researches. for example, Zhang et al. (2021)

adopted network analysis techniques to investigate on how covid-related topics evloved before and during the COVID-19 pandemic period. Also, Ebadi et al. proposed a method based on natural language processing and machine learning to help decision-makers to identify main covid-related topics (Ebadi et al., 2021).

Most of the existing topic evolution researches are based on topic modeling. Topic models are statistical approaches to discover the hidden semantic structure in a collection of documents (Blei & Ng, 2003; Blei, 2012). In topic modeling methods, each document is represented by multiple topics with different degrees of association (Belford & Greene, 2020). Topic models have found many applications in natural language processing (Jelodar et al., 2019). In many topic modeling methods, the article's textual content (and sometimes only its abstract) is used to model a topic. In other words, the texts of the articles may define the nature of a scientific area. For instance, word distribution can be considered in the articles' texts. Latent Dirichlet Allocation (LDA) (Hofmann, 2001; Blei & Ng, 2003) is one of the first topic modeling methods based on NLP techniques. It is also one of the most popular approaches in topic modeling (Dieng et al., 2020; Jung & Yoon, 2020). LDA adopts a bag of words approach to mine topics in the documents' corpus. Dynamic topic modeling (Blei & Lafferty, 2006) is also a well-known topic modeling approach, which focuses on word transitions of topics in fixed time-slots. Nowadays, Short texts published in social networks have created a good opportunity for researchers in the field of topic modeling. For example, Rashid et al. (2019) presented a fuzzy topic modeling (FTM) approach to model topics in large-scale short text documents. Topic modeling techniques can be applied to a specific field's lierature to examine how topic under that field evolve. For example, Kim et al. (2020) prposed a topic model to analyze trends on blockchain technology. They utilized Word2Vec and k-means methods to represent a context of a corpus and they examine treds annually.

Embedding approaches are among the main methods of extracting or modeling text topics. Methods such as Word2Vec (Le & Mikolov, 2014), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017) and probabilistic fastText (Bojanowski et al., 2017) are widely applied in this field. Similar to such methods, Wang et al. (2019) proposed a neural topic modeling approach based on Generative Adversarial Networks (GANs). They modeled topics by Dirichlet prior and used a generated network to achieve semantic patterns among latent topics.

The main aim of this paper is to investigate methods for quantifying convergence and divergence of research topics. In this context, topic evolution analysis methods can provide a tool for analyzing how two specific topics react to each other in different periods. For example, Gaul and Vincent (2017) formed topics as content-based document clusters and then detected the similarity between clusters in different periods. Furthermore, semantic changes of words over time are detected by Rudolph and Blei (2018) by performing word embedding and analyzing the semantic changes of word embedding in different periods of time. Additionally, the writers of this paper have provided a method to determine the convergence or divergence of research topics in artificial intelligence in earlier works (Harikandeh et al., 2021). It has been attempted to offer more thorough features of the benefits of the suggested technique in this article, as well as to display more thorough reports of the evolution of the topics, in addition to evaluating it in another domain (computer science).

Although many existing topic evolution analysis methods examine text content transitions (such as top words used in each topic), (Blei & Lafferty, 2006), the process of changing a topic can be identified without considering the content (text) under topics. For instance, co-citation analysis (Small, 1973) and co-word analysis (Callon et al., 1983) are utilized for topic evolution researches in recent years. By reviewing the existing related

works, it seems that the embedding approaches may result in efficient and simple methods for analyzing topic convergence and divergence patterns. Therefore, in this paper, we adopted a topic co-occurrence-based method in the articles.

Methodology

As discussed in the previous section, there are several approaches to model research topics. In this article, we embed each topic as a point in the multidimensional space. In this section, we review our proposed method for embedding research topics and how we use it to calculate the distance between the research topics. The proposed distance metric will measure the distance between research topics and their convergence and divergence over time.

Assumptions

We assume that there is a fixed and limited set of known research topics in computer science studies called "Field of Study" (FoS). For example, "Data Mining" and "Computer Security" are among FoSs under computer science. This paper aims to analyze the evolution of these FoSs in computer science studies. We also assume that we have access to a dataset of published scientific papers in computer science. We will analyze the papers in order to find patterns of topic evolution among the FoSs. For each paper in the dataset, we assume an already known set of research topics (FoS) related to the paper. For instance, a paper might be specified with different FoSs such as "data mining", "computer security", and "internet privacy". Nowadays, research articles may consider several different research topics. We assume that the higher the ratio of the involvement of two specific topics (FoSs) in the articles, the closer the two topics are. Therefore, we propose a simple model for computing the similarity of FoSs over time.

Quantified similarity among the research topics

Suppose that X is one of the n existing fields of study ($X \in \{FoS_1, FoS_2, ..., FoS_n\}$) and $A_i(X)$ is the set of all the papers that are related to X, published in year i. We consider the Jaccard similarity between two FoSs (X and Y) in the year i denoted by $Similarity_i(X, Y)$ as described in Eq. 1:

$$Similarity_i(X, Y) = \frac{|A_i(X) \cap A_i(Y)|}{|A_i(X) \cup A_i(Y)|} \tag{1}$$

Equation 1 represents a similarity metric which indicates how much two FoSs X and Y are involved in the same research articles published in the year i. Then, this similarity metric can be utilized for topic embedding. In order to examine the evolution of a topic (FoS) over time, we embed each of the n topics as a numeric vector. For each year i, the topic X is embedded as a vector named $\vec{V}_i(X)$ which is described in Eq. 2, in which $Y_1, ..., Y_n$ are the research topics (FoSs).

$$\vec{V}_i(X) = < Similarity_i(X, Y_1), Similarity_i(X, Y_2), ..., Similarity_i(X, Y_n) > \tag{2}$$

For any topic pair (X,Y), $0 \leq Similarity_i(X, Y) \leq 1$ and $Similarity_i(X, X) = 1$. As a result, for each year i we have n vectors of size n that are the embedded vectors of each topics in that year.

One of the aims of this article is to investigate the distance between topics and identify close, distant, converging, and diverging topics. Placing each topic in a vector space allows us to perform mathematical operations based on the embedded topic vectors. We utilize the Euclidean distance between embedded vectors as the distance metric for research topics. The Euclidean distance between topics X and Y is defined in Eq. 3, in which $\vec{V}_{i_k}(X)$ denotes the kth element in the $\vec{V}_i(X)$ vector:

$$Distance_i(X, Y) = |\vec{V}_i(X) - \vec{V}_i(Y)| = \sqrt{\sum_{k=1}^{n} (\vec{V}_{i_k}(X) - \vec{V}_{i_k}(Y))^2} \tag{3}$$

Now, we can measure the convergence or divergence of topics over different years with the aid of the presented distance formula in Eq. 3. A question may arise as to, despite the fact that we defined a Jaccard similarity metric in Eq. 1, why did we define the distance metric in Eq. 3 (We could calculate the distance between topics directly from Jaccard similarity). We will further examine that embedded-based distance (presented in Eq. 3) is more effective than jaccard-based distance in the context of analyzing topic evolution based on a case study in "Priority of Embedded Similarity over Jaccard Similarity"section.

Results

This section reports the results of studying topics evolution in computer science research in recent years. It is attempted to discover the hidden relationships between topics, particularly to understand how research topics move closer or farther apart based on their contribution to the published articles. We apply the proposed methodology described in "Methodology" section for analyzing the similarity of research topics over time.

Dataset

We employed the Aminer[1] version 13 dataset as the source for the published papers (Tang et al., 2008). This dataset contains information about over 5.3 million scientific papers published until the year 2021 and combines the Computer Science (CS) articles from Microsoft Academic Graph (MAG), the DBLP, and the ACM datasets. While the MAG dataset provides many of the attributes available in this dataset, the other collections supplement it and assure that the articles are CS-related. Also, this dataset is freely available and enables further research associated with the current article. Each paper may involve more than one "field of study" (FoS), and therefore, a set of FoSs is specified for each article in the dataset. The FoSs in this dataset are presented in a hierarchy of topics which are extracted from the Microsoft Academic Graph (MAG) project (Shen et al., 2018). MAG organizes topics into four distinct levels. Level zero contains general scientific topics such as computer science, mathematics, and engineering. Lower levels

[1] https://www.aminer.org/citation.

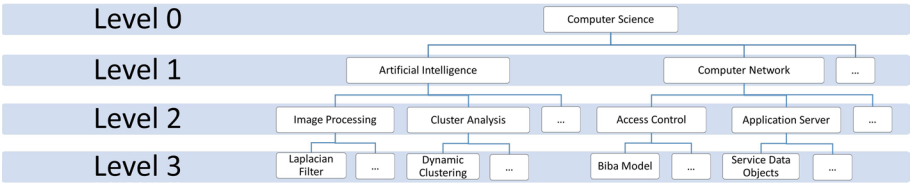


Fig. 1 An example of different levels of research topics (FoS) in the dataset

Table 1 The 34 computer science fields of study (FoS) and their assigned short-names

FoS	Short-name	Fos	Short-name
Artificial intelligence	ArtInt	Natural language processing	NLP
Machine learning	MachLrn	Software engineering	SoftEng
Data mining	DataMin	Computer security	ComSec
Knowledge management	KnoMng	Programming language	ProgLang
Distributed computing	DisCom	Simulation	Sim
Speech recognition	SpchRec	Pattern recognition	PattRec
Computer vision	ComVis	Embedded system	EmbSys
Multimedia	Mul	Computer engineering	ComEng
Information retrieval	InfRet	Library science	LibSci
Theoretical computer science	TCS	Operating system	OS
Database	DB	Computer hardware	ComHrd
Computer network	ComNet	Computer graphics (images)	ComGrp
Human computer interaction	HCI	Telecommunications	TeleCom
Real-time computing	RTC	Internet privacy	IntPri
Parallel computing	ParCom	Data science	DataSci
World Wide Web	WWW	Computer architecture	ComArch
Computational science	ComSci	Algorithm	Algo

in the hierarchy of research topics represent more specific fields of study. For instance, Computer Science has 34 FoSs in level 1, including Artificial Intelligence, Computer Network, and Computer Graphics. Figure 1 shows an example of the MAG hierarchy in different levels for the Computer science field of study. This paper focuses on the "Computer Science" field of study, which is a level-zero topic in the MAG hierarchy. For better visualization in the reports, we assigned a short name to each FoS. Table 1 lists sub-fields of Computer Science studies and their assigned short names. In this dataset, the information related to the papers published from 2019 to 2021 are not complete and stable yet. Therefore, we decided to consider only the papers published from 2000 to 2018 to maintain the produced reports' integrity. We also excluded the "Artificial Intelligence" (AI) FoS from our reported results since AI acts as the super-set of the four following FoSs in this dataset: Machine Learning, Pattern Recognition, Computer Vision, and Natural Language Processing (NLP). In other words, these FoSs altogether show the convergence/divergence of research topics concerning Artificial Intelligence. Since AI is the most trending FoS in this dataset, it biases all the forthcoming reports and is therefore excluded from our reports.

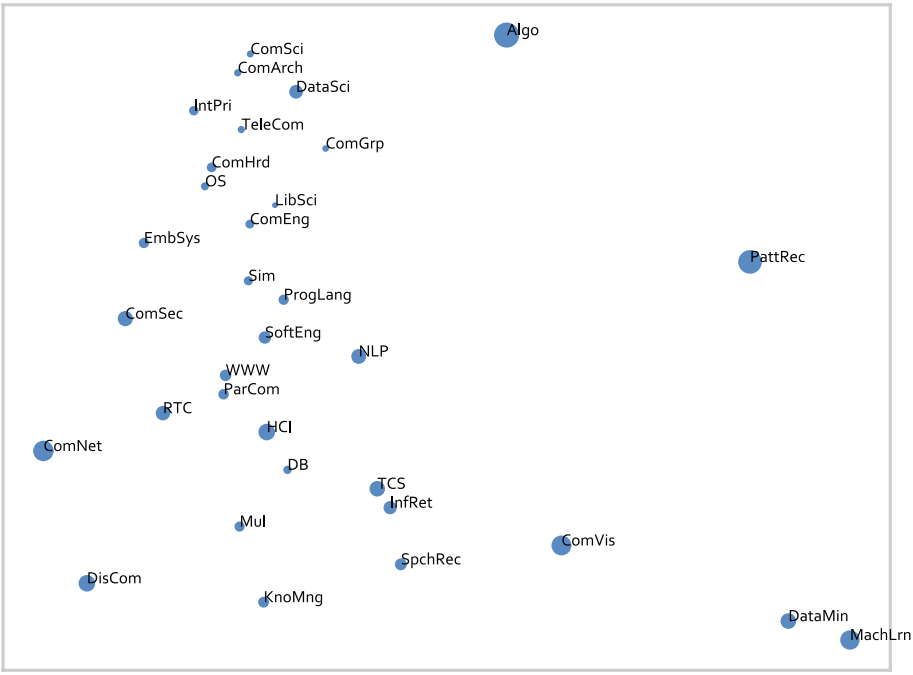


Fig. 2 Two-D plot of the research topics in 2018. Similar research topics are plotted closer in this visualization

Research topics visualization

In the first report, the condition of the research topics in a two-d plot is visualized in Fig. 2. As described in Eq. 2, each FoS in computer science studies is represented as a vector of 33 numeric elements (there are 33 FoSs in computer science in the dataset after excluding Artificial intelligence). Next, using the Principal Component Analysis (PCA) dimensionality reduction method (Jolliffe & Cadima, 2016), we reduced each FoS's vector to a two-dimensional vector in order to visualize them in a 2-D plot. Figure 2 describes the relative distance of topics in the year 2018. For example, as this plot shows, Machine Learning and Data Mining are close to each other, meaning that many published pieces of research in Data Mining are also related to Machine Learning. The node size also relates to the number of published papers regarding each FoS. We should also note that dimensionality reduction methods (such as PCA) cannot maintain all the relative topic distances in a low-dimensional space, and the 2-D plot is used to demonstrate a representation for describing the similarity of research topics. Therefore, more quantified and qualified results are reported in the following subsections to complement this plot's data.

Distant and close topics

One of the aims of this research is to find similar and distant research topics in computer science studies. Discovering the most adjacent and most distant FoSs in different years can give us a deeper insight into computer science topics. For instance, "Distributed

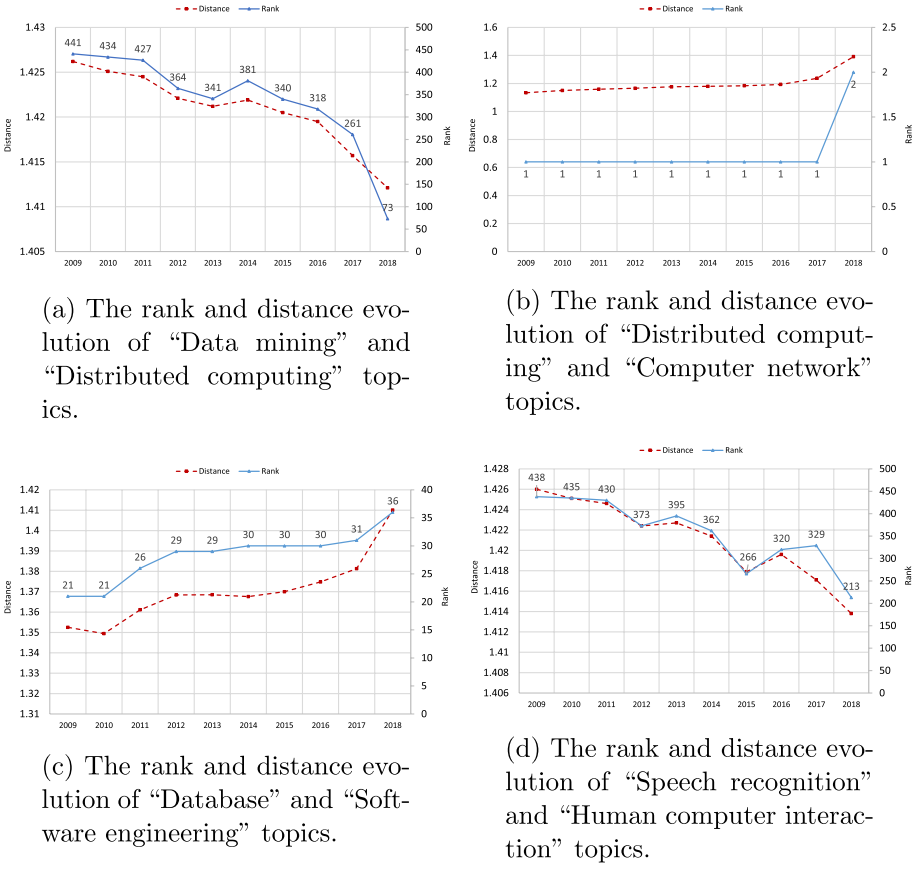
Table 2 Top 5 closest and top 5 furthest topics in 2018

Rank	1st FoS	2nd FoS
1	Machine learning	Data mining
2	Distributed computing	Computer network
3	Computer vision	Pattern recognition
4	Computer security	Internet privacy
5	Machine learning	Pattern recognition
.	.	.
524	Machine learning	Computational science
525	Computer vision	Computer network
526	Computer network	Pattern recognition
527	Distributed computing	Pattern recognition
528	Data mining	Embedded system

Computing" and "Computer Networks" were the most similar topic pair in computer science studies from 2000 to 2017 in the dataset, but "Machine learning" and "Data mining" became the most adjacent FoS pair in 2018, and we know that due to the large amount of information published in recent years, these two areas have multiple uses for extracting knowledge from unstructured data and are involved in solving numerous problems concurrently (Taranto-Vera et al., 2021). In order to provide a comprehensive representation for these research topics, we sorted the FoS pairs in computer science according to their pairwise distances, which resulted in a ranking of FoS pairs. Table 2 shows the top 5 most similar research topics, as well as the top 5 most distant FoSs in 2018. For instance, "Distributed Computing" is close to "Computer Networks" but far from the "Pattern Recognition" field. In this table, the first column shows the rank in the sorted list of FoS pairs in descending order, and the second and third columns display the FoS pair. For example, the first row has a rank equal to 1, indicating that "Machine Learning" and "Data Mining" are the most similar pair in the collection of all the FoS pairs. This pair makes sense since these two concepts have been becoming closer in recent years, as many researchers, such as Teng and Gong (2018), declared. On the other hand, the last row shows the rank of 528, implying that "Data Mining" and "Embedded Systems" have the least similarity among all FoS pairs. As reported in Richthammer et al. (2020), data mining methods have not yet been utilized much in embedded systems due to their lack of computational capacity and storage, justifying the results in the current study.

Convergence and divergence of research topics

One of the aims of this research is to establish a framework for examining how research topics converge or diverge concerning each other. For this purpose, we study the variations in the distance between research topics over time. We considered two metrics for the study of FoS pair evolution in each year under study. First, the similarity between every two topics, and second, the rank of the similarity among all the topic pairs are the metrics under consideration. As an example, Fig. 3 demonstrates the evolution of four sample topic pairs over 10 years. Figure 3a illustrates how "Data Mining" and "Distributed Computing" are becoming more similar over these years, according to their pair distance and pair distance rank (Fig. 3a). In the remainder of this article, topic pair rank will determine the rank of the FoS pair according to their pairwise distance among all the FoS pairs. For instance, Fig. 3a



(a) The rank and distance evolution of "Data mining" and "Distributed computing" topics.

(b) The rank and distance evolution of "Distributed computing" and "Computer network" topics.

(c) The rank and distance evolution of "Database" and "Software engineering" topics.

(d) The rank and distance evolution of "Speech recognition" and "Human computer interaction" topics.

Fig. 3 The changes in topic pair distance ranks (blue lines) and topic pair distances (red lines) from 2009 to 2018 (For better readability, the results of the last 10 years are displayed)

demonstrates that the topic pair rank of "Data Mining" and "Distributed Computing" has decreased from 441 in the year 2009 to 73 in 2018, indicating that these two topics have become more similar and related (covering more joint researches) throughout these years. The changes in the topic pair rank are more straightforward to interpret than the pair distance changes; therefore, we included pair ranks in our reports accordingly.

As another sample, Fig. 3b presents that "Distributed Computing" and "Computer Network" are very similar. However, their resemblance is slightly decreasing in the past years, and their distance is increasing. As Fig. 3c demonstrates, the similarity between "Database" and "Software Engineering" is slightly decreasing. Figure 3d also displays that "Speech Recognition" and "Human-Computer Interaction" are not similar topics. Though, they are moderately becoming more similar topics in recent years.

Followed by these intuitive examples, the most converging topics in recent periods are reported. In table 3, we investigate the pace of convergence between FoS pairs. In other words, regardless of how similar the two topics are, we seek to recognize how rapidly they are becoming more related. We considered two recent periods and analyzed the topic evolution in the two periods to satisfy this purpose. First, a 5-year period from 2009 to 2013, and

Table 3 The most converging research topic pairs from 2009 to 2018

FoS	$\text{Rank}_{[2009–2013]}$	$\text{Rank}_{[2014–2018]}$	$\Delta Rank$
Data mining – distributed computing	401.4	274.6	– 126.8
Speech recognition – human computer interaction	414.2	298	– 116.2
Distributed computing – computer engineering	328	223.4	– 104.6
Computer network – simulation	353.8	250.2	– 103.6
Theoretical computer science – Computer network	437.4	334.6	– 102.8
Database – computer network	411.6	309	– 102.6
Multimedia – natural language processing	468.2	368.6	– 99.6
Simulation – pattern recognition	419	323.4	– 95.6
Computer security – embedded system	281.6	191	– 90.6
Multimedia – computer security	357	269	– 88

The average rank of each FoS pair is shown in the second and third columns, which correspond to two time periods of [2009–2013] and [2014–2018]. In the forth column, the amount of change in the average distance rank of the topic pair is shown with respect to the two time intervals

second, another 5 years from 2014 to 2018 are assessed. Subsequently, we calculated the average distance rank for each FoS pair in each of the periods mentioned above. In the next step, we calculated the difference between the ranks in these periods, and which is called $\Delta Rank$ as defined Eq. 4. In this equation, $\overline{Rank(X,Y)}_{[i..j]}$ is the average distance rank of the FoS pair (X, Y) in the period from year i to j inclusive. Negative $\Delta Rank$ indicates the convergence of the two topics, while positive $\Delta Rank$ demonstrates their divergence. Therefore, the most converging topics are those with the smallest $\Delta Rank$ (the negative $\Delta Rank$ with the largest absolute value). Table 3 shows the FoSs that have had the most convergence in recent years. Although *Data mining* and *Distributed computing* are not very close (they stand about 274$^{\text{th}}$ place among 528 topic pairs), they are becoming more and more similar in recent years since the similarity rank is improved from 401 to 274. In other words, these two topics show the minimum $\Delta Rank$, which indicates they are moving towards each other, and therefore, more research that considers both of the topics is being published in recent years. A witness to the convergence of these two topics may be the development of a field named as "distributed data mining," which deals with extracting data from various information sources (Zeng et al., 2012). The same interpretation may also be considered for the other rows of Table 3.

$$\Delta Rank(X,Y) = \overline{Rank(X,Y)_{[2014–2018]}} - \overline{Rank(X,Y)_{[2009–2013]}} \tag{4}$$

Topic trends

So far, we have examined the evolution of each pair of FoSs according to their pairwise distances and convergence/divergence patterns over the years. Nevertheless, it is also possible to consider the amount of change in every FoS in recent years separately in topic evolution studies. In this regard, Fig. 4 shows the topic changes in recent years. For instance, the figure shows that the "Pattern Recognition" topic has experienced the most variation among the Computer Science research topics in recent years. In this figure, colors indicate the amount of change in a research topic (FoS) throughout recent years. First, as Eq. 5

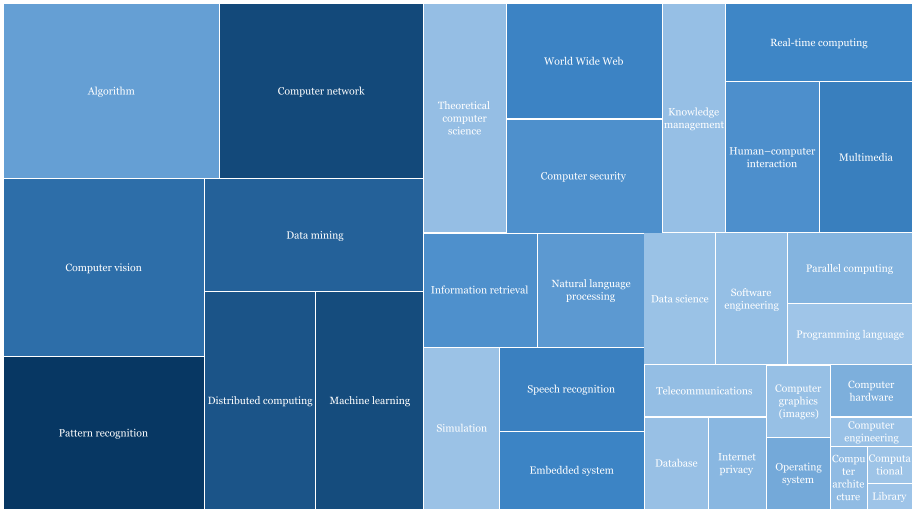


Fig. 4 Treemap of most changed FoSs. The size of block denotes the count of papers related to corresponding FoS from 2009 to 2018 and the color shows the amount of change oa each FoS such that deeper color representing more change

presents, we consider the average embedded vector corresponding to each FoS X over the years i to j as $\vec{V}_{[i..j]}(X)$. Then, as Eq. 6 shows, we define $Diff(X)$ as the amount of difference (movement) of the embedded vector corresponding to the topic X in recent years. The studied period is from the 5 years of [2009–2013] to the 5 years of [2014–2018]. In Fig. 4, the colors show $|Diff(X)|$ for different research topics where darker colors indicate a larger amount of changes.

$$\vec{V}_{[i..j]}(X) = \frac{\sum_{y=i}^{j} \vec{V}_y(X)}{j - i + 1} \tag{5}$$

$$Diff(X) = \left| \vec{V}_{[2014-2018]}(X) - \vec{V}_{[2009-2013]}(X) \right| \tag{6}$$

Priority of embedded similarity over jaccard similarity

Here, we can further investigate the priority and effectiveness of the embedded-based distance metric ($Distance_i(X, Y)$ as defined in Eq. 3) over the Jaccard-based distance metric ($1 - Similarity_i(X, Y)$. $Similarity_i(X, Y)$ is defined in Eq. 1) in the context of research topics evolution. Some examples will be reviewed on the utilized dataset to support this analysis. As an intuitive example, consider three research topics X, Y, and Z as: X="Computer architecture", Y="Telecommunications", and Z="Machine learning" in the year 2018 (i=2018). Based on the utilized dataset, we recognize that $Similarity_i(X, Y) = 0$ and $Similarity_i(X, Z) = 0$, but there is informative embedded-based distances between the FoS pairs <X,Y> and <X,Z> since $Distance_i(X, Y) < Distance_i(X, Z)$. In other words, although published articles related to "Computer Architecture" had neither"Telecommunication"

Table 4 Most Similar Topics to Telecommunication, Computer Architecture, and Machine Learning

Telecommunications	Computer Architecture	Machine Learning
Computer network	Embedded system	Data mining
Library science	Parallel computing	Pattern recognition
Real-time computing	**Computer engineering**	Algorithm
Computer engineering	Computer hardware	Theoretical computer science
Computer security	**Operating system**	Natural language processing
Algorithm	**Real-time computing**	Information retrieval
Operating system	**Simulation**	Computer vision
Simulation	*Multimedia*	Data science
Distributed computing	**Distributed computing**	Speech recognition
Multimedia	**Computer network**	*Multimedia*

The pair $< Telecommunication, ComputerArchitecture >$ has more common topics (topics with bold texts) than the pair $< ComputerArchitecture, MachineLearning >$ (italics topics)

nor "Machine Learning" tags in the year 2018, according to the $Distance_i(X, Y)$ metric, "Computer Architecture" is more similar to "Telecommunication" than "Machine Learning". However, is this similarity meaningful? To answer this question, we consider ten closest topics to each of these three topics, based on the Jaccard-similarity definition ($Similarity_i(X, Y)$ as defined in Eq. 1)). Table 4 shows the most similar topics to each of "Telecommunications", "Computer architecture", and "Machine learning" subjects. By reviewing this table, it is understood that "Computer Architecture" has seven common similar topics to "Telecommunication", while it has only one similar topic to "Machine Learning". Therefore, it is reasonable to consider "Computer Architecture" more similar to "Telecommunication" than to "Machine Learning". This case was an intuitive example of the priority of embedded-based distance over the Jaccard-based distance.

As another intuitive case, we consider the trend of distance changes between two sample research topics, "Database" and "Computer Hardware", which is illustrated in Fig. 5. In this figure, the trend of changes is seen both in terms of the Jaccard-based and the embedding-based distance. The changes in the Jaccard-based method display many fluctuations, while in the embedding method, these changes are smoother and more constant. This is because the state of other FoSs contributes to the definition of the embedding-based distance metric, keeping it from sudden inappropriate changes.

Conclusion

This paper established a framework for investigating the evolution of research topics and their convergence and divergence patterns. To this purpose, we proposed a method for embedding each research topic in a numerical vector representation. Then, we proposed a distance metric to calculate the amount of dissimilarity among research topics. Based on this methodology, we utilized a dataset of published papers in the field of computer science, and we performed a case study on research topics evolution in this area. We presented comprehensive reports to illustrate the evolution of different computer science subtopics in recent years.

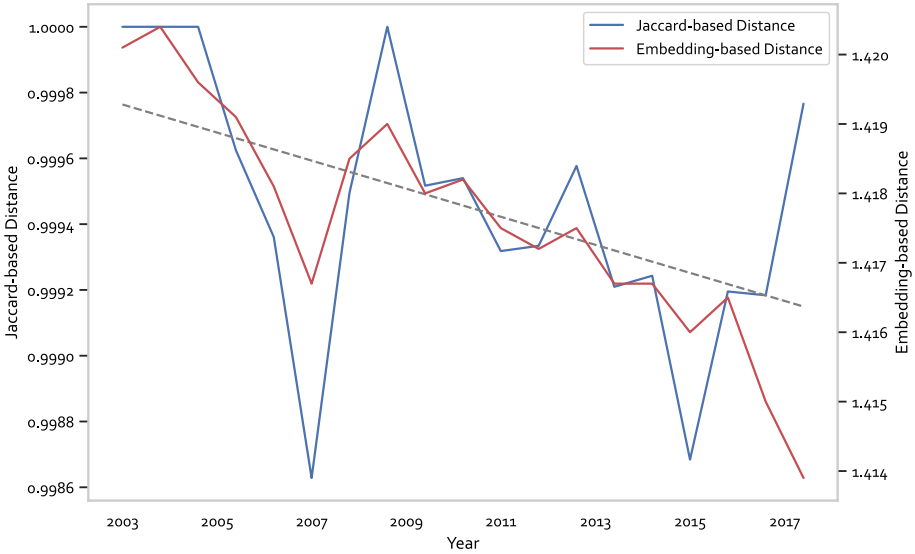


Fig. 5 Comparison of embedding-based and jaccard-based distance metrics for a sample FoS pair: "Database" and "Computer Hardware". The embedding-based approach shows smoother changes

Taking the same approach, we will concentrate on some other problems in our future research works. It is possible to investigate the application of the proposed methodology on different research areas other than computer science. Moreover, proposing a method to predict the topic evolution in the future based on machine learning techniques can be beneficial for the research community to better anticipate the direction of topics in the years to come. Topic evolution prediction has many potential applications in research management and funding because it may give researchers a more precise view of the future of their research fields and the impact of other topics in the future. The methodology presented in this article can also be applied in other research areas. For instance, we can utilize this method to analyze social media and social network contents instead of research articles. Notably, how a social network's hashtags (such as Twitter or Instagram hashtags) evolve over time may be studied.

References

Alam, M. H., Ryu, W.-J., & Lee, S. (2017). Hashtag-based topic evolution in social media. *World Wide Web, 20*(6), 1527–1549.

Allan, J. (2002). *Introduction to topic detection and tracking* (pp. 1–16). Boston: Springer.

Becker, H., Naaman, M., & Gravano, L. (2009) Event identification in social media, in *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*.

Belford, M., & Greene, D. (2020). Ensemble topic modeling using weighted term co-associations. *Expert Systems with Applications, 161*, 113709.

Blei, D. M. (2012). *Probabilistic topic models* (Vol. 55, pp. 77–84). New York, NY: Association for Computing Machinery.

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

Blei, D. M., & Lafferty, J. D. (2006) Dynamic topic models, in *Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery*, New York, NY, USA, pp. 113–120.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Callon, M., Courtial, J.-P., Turner, W., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information Sur Les Sciences Sociales - SOC SCI INFORM, 22*, 191–235.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*, 1175–1189.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.

Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., & Wong, A. (2021). Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics, 126*(1), 725–739.

Evans, J., & Rzhetsky, A. (2011). Advancing science through mining libraries, ontologies, and communities. *The Journal of Biological Chemistry, 286*, 23659–23666.

Fernández-Isabel, A., Barriuso, A. A., Cabezas, J., Martín de Diego, I., & Viseu Pinheiro, J. J. (2020). Knowledge-based framework for estimating the relevance of scientific articles. *Expert Systems with Applications, 161*, 113692.

Gaul, W., & Vincent, D. (2017). Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification, 11*(1), 159–178.

Guo, W., Caliskan, & A. (2021) Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery*, New York, NY, USA, (pp. 122–133).

Harikandeh, S. R. T., Aliakbary, S., & Taheri, S. (2021) Towards study of research topics evolution in artificial intelligence based on topic embedding, in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*. (pp. 406–411).

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, CIKM '09, Association for Computing Machinery*, New York, NY, USA, (pp. 957–966).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1), 177–196.

Huang, C., Yang, C., Wang, S., Wu, W., Su, J., & Liang, C. (2020). Evolution of topics in education research: A systematic review using bibliometric analysis. *Educational Review, 72*(3), 281–297.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169–15211.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical. *Physical and Engineering Sciences, 374*(2065), 20150202.

Jung, S., & Yoon, W. C. (2020). An alternative topic model based on common interest authors for topic evolution analysis. *Journal of Informetrics, 14*(3), 101040.

Kalyanam, J., Mantrach, A., Saez-Trumper, D., & Vahabi, H., Lanckriet, G. (2015) Leveraging social context for modeling topic evolution, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, New York, NY, USA, (pp. 517–526).

Kay, L., Newman, N., Youtie, J., Porter, A. L., & Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance, Journal of the Association for. *Information Science and Technology, 65*(12), 2432–2443.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications, 152*, 113401.

Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences, 117*(4), 1910–1916.

Le, Q., & Mikolov, T. (2014) Distributed representations of sentences and documents II (pp. 1188–1196).

Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), 102611.

Masood, M. A., & Abbasi, R. A. (2021). Using graph embedding and machine learning to identify rebels on twitter. *Journal of Informetrics, 15*(1), 101121.
Pennington, J., Socher, R., Manning, C. (2014) GloVe: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, Doha, Qatar, (pp. 1532–1543).
Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics, 14*(3), 101047.
Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management, 56*(6), 102060.
Richthammer, V., Scheinert, T., & Glaß, M. (2020) Data mining in system-level design space exploration of embedded systems. (pp. 52–66).
Rudolph, M., & Blei, D. (2018) Dynamic embeddings for language evolution. (pp. 1003–1011).
Sayyadi, H., Hurst, M., & Maykov, A. (2009) Event detection and tracking in social streams, in E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, B. L. Tseng (Eds.), *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California USA, May 17-20, 2009*. The AAAI Press.
Shen, Z., Ma, H., & Wang, K. (2018) A web-scale system for scientific knowledge exploration, in *Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics*, Melbourne, Australia, pp. (87–92).
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*, 265–269.
Song, M., Heo, G., & Kim, S. (2014). Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in dblp. *Scientometrics, 101*, 397–428.
Taheri, S., & Aliakbary, S. (2022). Research trend prediction in computer science publications: A deep neural network approach. *Scientometrics, 127*(2), 849–69.
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 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, New York, NY, USA, 2008*. pp. (990–998).
Taranto-Vera, G., Galindo-Villardón, P., Merchán-Sánchez-Jara, J., Salazar-Pozo, J., Moreno-Salazar, A., & Salazar-Villalva, V. (2021). Algorithms and software for data mining and machine learning: A critical comparative view from a systematic review of the literature. *The Journal of Supercomputing, 77*(10), 11481–11513.
Teng, X., & Gong, Y. (2018). Research on application of machine learning in data mining. *IOP Conference Series: Materials Science and Engineering, 392*(6), 062202.
Van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics, 38*(1), 205–218.
Wang, R., Zhou, D., & He, Y. (2019). Atm: Adversarial-neural topic model. *Information Processing & Management, 56*(6), 102098.
Xia, H., & Liu, P. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics, 103*, 101–134.
Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. *Scientometrics, 117*(2), 973–995.
Zeng, L., Li, L., Duan, L., Lu, K., Shi, Z., Wang, M., Wu, W., & Luo, P. (2012). Distributed data mining: A survey. *Information Technology and Management, 13*(4), 403–409.
Zhang, Y., Cai, X., Fry, C. V., Wu, M., & Wagner, C. S. (2021). Topic evolution, disruption and resilience in early COVID-19 research. *Scientometrics, 126*(5), 4225–4253.