Unraveling Scientific Evolutionary Paths: An Embedding-Based Topic Analysis

Article *in* IEEE Transactions on Engineering Management · October 2023

DOI: 10.1109/TEM.2023.3312923

5 authors, including:

Qianqian Jin
Beijing Institute of Technology
6 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Hongshu Chen
Beijing Institute of Technology
40 PUBLICATIONS **820** CITATIONS
SEE PROFILE

Yi Zhang
University of Technology Sydney
126 PUBLICATIONS **2,546** CITATIONS
SEE PROFILE

Xuefeng Wang
Beijing Institute of Technology
75 PUBLICATIONS **999** CITATIONS
SEE PROFILE

Unraveling Scientific Evolutionary Paths: An Embedding-based Topic Analysis

Qianqian Jin[1], Hongshu Chen[1*], Yi Zhang[2], Xuefeng Wang[1], Donghua Zhu[1]

[1] School of Management and Economics, Beijing Institute of Technology, China
[2] Australian Artificial Intelligence Institute and the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

Abstract: Understanding the evolution of knowledge has been and will continue to be the key task of science, technology, and innovation management. Existing research on evolutionary path identification relies primarily on traditional co-occurrence analysis and bag-of-words-based (BOW) models for topic extraction. However, these approaches have limitations in effectively capturing the underlying semantics and linkages of the topics. In this paper, we propose a novel embedding-based methodology for scientific evolution analysis, in which word embedding, document embedding, clustering, and network analysis are applied to extract topics, measure topical semantic similarities, and quantitatively distinguish topics' evolutionary states. We first perform benchmark experiments to demonstrate that Doc2vec generally outperforms the BOW-based models in topic extraction prior to evolution analysis. We then consider topic consistency in vector spaces to identify evolutionary states including newborn, convergence, inheritance, and extinction. Scientific evolutionary paths are finally unraveled based on topic similarity matrixes and evolutionary states. We conduct a case study on object detection research to validate the effectiveness of our methodology. The empirical results, validated by domain experts, demonstrate that the proposed methodology is capable of effectively revealing patterns of knowledge inheritance and integration. Consequently, this methodology can be used to improve decision-making processes in future innovation management.

Keywords: evolutionary paths; evolution analysis; topic extraction; embedding; doc2vec; word2vec

Please cite as:
Jin, Q., Chen, H., Zhang, Y., Wang, X., & Zhu, D. (2023). Unraveling Scientific Evolutionary Paths: An Embedding-Based Topic Analysis. IEEE Transactions on Engineering Management, 1-15. doi:10.1109/TEM.2023.3312923.

* Hongshu.Chen@bit.edu.cn

1. Introduction

As science is fast-moving, understanding how scientific knowledge accumulates, inherits, and evolves has become and will continue to be the key task of science, technology, and innovation (ST&I) studies [1]. Scientific evolution analysis – being considered as much an art as a science – unravels the nature and patterns of knowledge flow in the evolutionary paths, enables better plans and decisions in practice, facilitates knowledge integration and recombination, and consequently, accelerates the advance of ST&I [2, 3]. From the perspective of innovation recombination theory, emerging ideas and topics continue to connect with the existing ones, expanding the boundary of scientific knowledge, and heralding the future trend [4]. The dynamic detection of relations between emerging and existing topics over different periods thus becomes one of the main issues of evolutionary path identification [5].

Topic extraction and connection from massive scientific documents are the foundation of scientific evolution analysis [6]. The previous research on evolution analysis, which mainly built on bibliometrics-based topic extraction, has yielded significant outcomes, and become the cornerstone of further studies on evolutionary path recognition [7-9]. The continuous growth of research papers, patents, and other scholarly artifacts results in an exponential increase in the number of features involved, such as words and phrases, advanced feature representation methods combined with increased computational power begin to provide new inspirations for topic extraction and further evolution analysis [10]. In the last decade, the Term Frequency-Inverse Document Frequency model (TF-IDF) combined with clustering algorithms, as well as the Latent Dirichlet Allocation (LDA) model, have emerged as the most commonly employed topic extraction tools for investigating scientific evolution based on textual data [11-13]. However, both TF-IDF and LDA are premised on the Bag-of-words (BOW) assumption. These existing methods face challenges in effectively capturing the inner semantics of topics and are also hindered by the high-dimensional problem [14, 15]. Consequently, they may provide inaccurate and inefficient results when attempting to extract topics and identify their connections using similarity measurements.

In recent years, word and document embedding algorithms have caught the attention for their promising capability to capture token-level syntactic and semantic information from contexts via numeric vector mapping [16-19]. They produce fixed-length vectors, which are convenient for similarity measurement, thus holding great potential for discovering semantic relations and mining potential evolutionary patterns [20]. Despite the current popularity of embedding techniques, the field of embedding-based topic extraction and evolution analysis is still in its early stages of development [21]. The question of whether embedding methods outperform conventional topic representation approaches and, more significantly, if they are ideal for unraveling scientific evolutionary paths, remain unanswered, and rigorous comparison and investigation are needed [22, 23].

In addition to efficient topic extraction and connection, how to reveal evolutionary patterns is another key subject that has received much attention. Identifying the evolutionary states of topics is a significant but challenging task, often relying on methods such as setting similarity thresholds in the vector space or relying on expert judgment [13, 24]. Irrespective of whether BOW-based or embedding-based methods are employed for evolution analysis, differentiating between topics that have genuinely evolved and those that are simply highly semantically related remains a challenging

task [6]. The question of how to depict the knowledge flow by linking topics across different periods with full consideration of their consistency and dynamic changes remains unresolved [10]. In addition, there is insufficient attention paid to topic designation in existing research, although it is important and necessary before identifying topic evolution.

Facing these challenges, this paper aims to provide a methodology for unraveling scientific evolutionary paths with embedding techniques. To address the gap in showcasing the effectiveness of embedding methods in capturing semantics and extracting topics prior to evolutionary path identification, we conducted a comparative analysis. We evaluated the performance of word2vec, doc2vec, BERT, and SciBERT against the baseline results of TF-IDF and LDA. Due to its superior performance, we selected Doc2vec in combination with k-means clustering to conduct a further scientific evolution analysis. In addition to creating an embedding-based document-level vector space, we employed word2vec to generate a word-level vector space. Multiple networks were constructed to identify significant terms for topic representation and designation. Considering both topics' similarities and consistency in vector spaces, four evolutionary states are identified, including newborn, knowledge convergence, knowledge inheritance, and extinction, using a "head-body" approach. Finally, to demonstrate the feasibility and effectiveness of our methodology in unraveling scientific evolutionary paths, a case study on object detection research is conducted with validation.

The remainder of this paper is organized as follows: Section 2 reviews mainstream methods for scientific evolution analysis and relevant studies on evolutionary state identification, and summarizes several typical embedding approaches in topic extraction and analysis. Section 3 describes the two procedures of the proposed methodology – topic extraction and scientific evolutionary path unraveling – after demonstrating the effectiveness of embedding approaches. Then we conducted an empirical case study on the field of object detection in Section 4, where we also discussed and validated the results. Finally, Section 5 concludes with the contributions, limitations, and prospects of this research[†].

2. Literature review

2.1. Scientific evolution analysis

Scientific evolutionary paths of a field can uncover its knowledge flow patterns, unearth emerging research frontiers, and forecast future trends [3, 25]. Extant research on scientific evolution analysis is generally conducted from two perspectives [5, 10]. One perspective to capturing the evolutionary trajectories of science involves applying main path analysis (MAP) on citation networks, focusing on identifying the main paths of knowledge flow within the network. Another perspective involves extracting topics and tracking their development over time using bibliometric methods or text-mining techniques [9].

2.1.1. Main path analysis in citation networks

Citations are widely recognized as a proxy for knowledge flow and diffusion [26]. In Garfield's historiography map, new ideas and discoveries are dependent upon the prior ones [27]. Building on citation networks, the main path analysis (MAP) method, which traces the most significant paths

[†] The data and codes of this paper are available on GitHub. https://github.com/ycsun1972/ScientificEvolution

according to the connectivity of citation links, has been widely used to detect scientific evolution trajectories [5, 28]. For example, Huang, et al. [29] applied global and local MAPs to identify technological trends of dye-sensitized solar cells. Chen, et al. [30] proposed a semantic-enhanced MPA method to simultaneously extract multiple evolutionary paths in the field of lithium-ion battery. With the aim of lessening network complexity and fully capturing important patents, Park and Magee [31] put forth a knowledge persistence-based MAP approach to simultaneously search backward and forward paths from high-persistence patents.

2.1.2. Topic extraction and evolutionary paths mapping

Topics in scientific publications can be represented by clusters detected in co-word, co-citation, or bibliographic coupling networks, using visualization tools or community detection techniques such as Girvan and Newman's division algorithm [32, 33]. Through depicting changes in these topics over time, the evolutionary path of science can be demonstrated [34]. In this way, Katsurai and Ono [7] constructed multiple co-word networks over years, in order to identify domains' emerging trends from a dynamic perspective. Majdouline, et al. [35] used network attributes to describe the temporal evolution of co-citation networks in technological entrepreneurship research. Mariani and Borghi [8] applied bibliographic coupling and network analysis to explore the knowledge structure and evolution trend of Industry 4.0.

Facing the explosive growth of scientific textual data and advancement of text mining techniques, document representation methods such as TF-IDF and LDA, begin to serve as important tools for turning scientific literature into comprehensible knowledge, and have received widespread acceptance from researchers who are devoted to scientific evolution analysis [4, 12, 13]. For example, Yang, et al. [13] used the Lingo method to extract topics and revealed the topic evolution of interindustry technology linkages. Zhang, et al. [36] combined TF-IDF and K-means to cluster terms into topics, and created a chronological roadmap to capture shifting topical emphasis in big data research. Song and Suh [37] applied topic modeling and network analysis to explore technological emergence and integration in industrial safety patents.

Despite the fact that TF-IDF and LDA have provided rich results in topic extraction and scientific evolutionary analysis, they have shortcomings. TF-IDF is a one-hot-encoding-based method; it only takes numerical statistics of words into consideration, ignoring semantic relations between words and context [14]. Although LDA can reveal global relations between documents and words, it currently fails to capture token-level syntactic and semantic information from contexts [38]. Moreover, TF-IDF and LDA are premised on the bag-of-words (BOW) assumption, perceiving that documents with similar composition and proportion of words could express the same topic [22]. The dimension of sparse vectors that bag-of-words-based approaches generate can increase in an incredible way with the growth of the corpus, which is an intrinsic limitation. Therefore, there is still much room to further improve the existing methods.

2.2. Embedding-based topic extraction and analysis

In recent years, neural network-based embedding approaches shine in the limelight for scientific text mining [39, 40]. As they are capable of mapping words or sentences into fixed-length numeric vectors, latent semantics in massive textual data can be translated into low-dimensional dense space [41]. Word2vec is one of the most representative embedding techniques using neural networks [16].

Building on the main idea of word2vec, doc2vec considers the order of words within sentences, goes beyond the word level, and makes it possible to acquire distributed vector representation at a document level [17]. These two models are based on static embedding that does not change with context once learned. More recently, the Bidirectional Encoder Representations from Transformers (BERT) model, which depends on dynamic embedding and learns new word vectors from the corpus based on the context, has caught the attention of researchers for topic extraction purposes. As derivatives of BERT, SciBERT and BioBERT focuses primarily on scientific and biomedical publications, respectively, and have already been shown to be very effective in scientific and biomedical text-mining tasks [42, 43]. Compared to static embedding, dynamic embedding might better capture the meaning of words in a context environment, however, may require more training time and cost [44].

Word2vec, BERT, and their derivative approaches have demonstrated their efficiency in turning massive textual data into dense vectors [18, 19], yet further clustering and learning are still needed for topic extraction and representation. Text clustering emphasizes semantic connections and statistical properties between documents [36]. Based on classic and common clustering methods such as k-means, DBSCAN, and spectral clustering, some achievements and insights have been acquired in topic analysis research [45]. For example, Angelov [46] integrated doc2vec and HDBSCAN to propose a Top2vec method and found it can find more representative and informative topics than topic models such as LDA and PLSA. Building on the BERTopic method proposed by Grootendorst [21], Sanchez-Franco and Rey-Moreno [47] clustered BERT tensors to extract topics from reviews of the Airbnb website, to investigate the relations between users preferences and locations. Zankadi, et al. [48] used BERTopic to extract topics from the comments of MOOC learners and identify their topical interests in social media.

Although embedding approaches combined with clustering methods have been used for topic extraction and analysis, no consensus has been reached on whether embedding methods are superior to classic topic representation techniques and, more importantly, whether they are appropriate for evolutionary state identification. Another concern lies in how to select terms from clustered documents to represent and designate topics [21, 46]. Even though there has been substantial development, embedding-based topic extraction is still in its infancy. More attempts are needed to validate and enrich existing research.

2.3. Evolutionary states identification for linked topics

To unravel the dynamic evolutionary processes of a field, extracted topics should be linked for scientific evolutionary state identification [25, 49]. Earlier research – which reveals evolutionary relationships relying solely on expert knowledge and experience – can be inaccurate, inefficient, and costly with the surge of scientific publications [24]. Much of the current literature links topics and determines their evolutionary states by assessing topics' commonalities, such as semantic similarity, structural similarity, and so forth [5].

Based on calculating semantic similarities of topics, Xie, et al. [2] mapped scientific development of topics extracted from multilingual publications. Zhang, et al. [10] set two thresholds to divide topics into four categories: novel, evolved, dead, and death and resurgence. Yang, et al. [13] distinguished five evolutionary states and examined topics' stability, heredity, and variability according to the knowledge gene theory.

In more recent studies, various topical features are being considered and integrated to identify evolutionary routes. For example, Wu, et al. [50] determined the evolutionary status of research themes based on their novelty, structure, and attention level. Xu, et al. [5] assessed the degree of topic association by considering the strength of topic linkages in multiple networks such as co-word and co-author networks. Miao, et al. [24] produced a semantic analysis method to extract terms involving products, functions, and technologies from patents, and constructed the technology road mapping according to the term structure and opinions of domain experts.

Despite considerable efforts and fruitful outcomes, identifying the evolutionary states of topics remains a challenging task. It is difficult to distinguish whether they have "evolved" or are merely semantically connected [51]. Therefore, further effort is needed.

3. Embedding-based topic extraction for scientific evolution analysis

To establish a solid foundation for scientific evolution analysis, this study is to initially investigate both mainstream and emerging embedding methods, including word2vec, doc2vec, and BERT, and compare their superiority in capturing semantics and extracting topics via several benchmarks. The optimal word-level and document-level embedding methods would be selected for topic extraction and scientific evolutionary path identification.

Word2vec is the well-known prominent neural network embedding method. Using the CBOW model and negative sampling technique to train word2vec, each unique term would be represented as a γ-dimensional vector. Doc2vec inherits the same main idea as word2vec. For each document, the DBOW model this research uses would return a δ-dimensional vector[‡]. As for the most cutting-edge embedding model BERT, its transformer layers can model the dependence between tokens and, in doing so, capture semantic and syntactic information from context. The hidden-states of the last transformer layer are used to represent feature vectors of corresponding documents. Then each document will be represented as a 768-dimensional vector[§].

[‡] This research uses the Python Gensim toolkit to train word2vec and doc2vec models. (https://radimrehurek.com/gensim/)

[§] This research chooses the pretrained BERT model bert-base-uncased with 12-layer, 768-hidden, and 12-heads to generate document vectors (https://huggingface.co/bert-base-uncased). As for the SciBERT and BioBERT, the scibert-scivocab-uncased developed by Allen Institute for AI, and the biobert-base-cased developed by DMIS-Lab, are used (https://github.com/allenai/scibert/; https://github.com/dmis-lab/biobert).

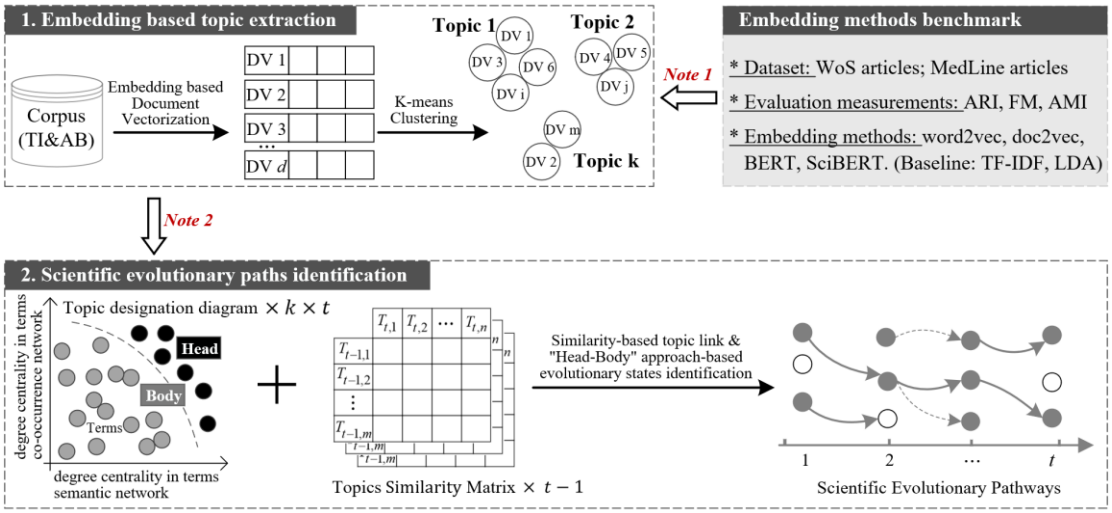

Fig. 1 Framework of embedding-based topic extraction for scientific evolution analysis.

Note 1: The effectiveness of candidate embedding methods is benchmarked, then only the optimal word-level and document-level methods are selected for further scientific evolution analysis.

Note 2: When performing scientific evolution analysis, k-means clustering should be performed multiple times for each year's documents. This complete process, as well as parameter setting for doc2vec and k-means, will be described in more detail in Section 3.2. Furthermore, Section 3.3 demonstrates the process of acquiring each topic's designation diagrams and topics' similarity matrices.

3.1. Embedding method benchmark for topic extraction

In order to select an effective and efficient method for unraveling scientific evolutionary paths, we compared the performance of aforementioned embedding approaches working in capturing content semantics and extracting research topics. Here, document clustering tasks with the k-means algorithm were performed. For word2vec, to go beyond word level and achieve document level representations, we employed an average, as well as a TF-IDF weighted average of all words in the document. BOW-based models, including TF-IDF and LDA, were chosen as baselines, as the benefits of using these two approaches for topic analysis have been widely approved.

Two bibliometric datasets with predefined taxonomies – Web of Science (WoS) and MedLine – were utilized to gather publications for model benchmark. WoS is a multidisciplinary classification system that divides research areas into five main classes and more than 250 subdivided Categories (WCs) [52]. We selected 10 diverse WCs that contain both fundamental and applied disciplines. For each category, 5000 articles were randomly retrieved on December 5, 2021. MedLine is one of the most representative databases in the field of life sciences, within which a hierarchically organized vocabulary, known as the Medical Subject Headings (MeSH) thesaurus, is used to index and catalog scientific publications. We collected articles affiliated with 10 different Leaf Mesh, which are derived from their respective root MeSH. For each subject, we downloaded 5000 articles on December 6, 2021. The descriptions of two datasets can be found in Appendix B.

We considered both precision and energy efficiency to compare the performance of candidate embedding methods with the baseline models of TF-IDF and LDA on clustering tasks for the two prepared datasets. If the embedding methods can properly capture semantics of textual data,

documents with high similarity are supposed to group together. Three metrics, including ARI, FM, and AMI, were calculated based on formulas provided in Appendix B [23]. The mean values of these indicators are shown in Tables 1 and 2, which generally imply that embedding methods outperform BOW-based models in terms of semantic encoding. The optimal performance and minimal time consumption are highlighted in bold. The time spent on document vectorization and clustering for each model was also recorded to assess the energy efficiency of candidate methods[**]. To minimize the impact of hyperparameters on overall performance, we test all models with different sets of hyperparameter settings. The final parameter settings are also listed in Appendix B, which might serve as guidance for future studies where there is a lack of ground-truth topic labels.

Table 1 Comparison of document representations on the WoS dataset.

	Document representation methods	ARI	FM	AMI	Training Time
Bag-of-words-based model	TF-IDF	0.6256	0.6672	0.7061	26m 49s
	LDA	0.3700	0.4715	0.5427	15m 21s
Embedding based model	word2vec (unweighted)	0.6931	0.7241	0.7333	24m 16s
	word2vec (TF-IDF weighted)	0.5741	0.6189	0.6673	26m 44s
	doc2vec	0.7733	0.7960	0.7896	**2m 39s**
	BERT	0.7009	0.7310	0.7461	653m 24s
	SciBERT	**0.7884**	**0.8097**	**0.8026**	654m 42s

Table 2 Comparison of document representations on the MedLine dataset.

	Document representation methods	ARI	FM	AMI	Training Time
Bag-of-words-based models	TF-IDF	0.3518	0.4465	**0.5861**	24m 30s
	LDA	0.2610	0.3406	0.3664	18m 19s
Embedding-based models	word2vec (unweighted)	0.3218	0.3912	0.4396	25m 18s
	word2vec (TF-IDF weighted)	0.3659	0.4312	0.4788	23m 45s
	doc2vec	**0.4123**	**0.4718**	0.5092	**3m 31s**
	BERT	0.2983	0.3704	0.4198	649m 59s
	SciBERT	0.3711	0.4364	0.4871	650m 29s
	BioBERT	0.3824	0.4469	0.5046	648m 04s

Some insights can be summarized from Tables 1 and 2. Firstly, word2vec performs slightly better than the BOW-based models. At the same time, it cannot be asserted that the weighted word2vec is superior to the unweighted one. Secondly, doc2vec outperforms most models in this clustering task. This finding is in agreement with studies conducted by Dai, et al. [53], which claimed that doc2vec is better than LDA and can get useful semantic results. As Lau and Baldwin [54] have pointed out, doc2vec is superior, especially when dealing with long documents. In addition, doc2vec is the most energy-efficient model compared to others, taking only about three minutes to train on the corpus

[**] We used two CPUs with 24 cores to train all the models.

that contains 50000 documents. Thirdly, BERT and its derivatives also show excellent performance. In particular, SciBERT is superior to others when dealing with scientific publications, while BioBERT is more suitable for biomedical text mining. However, BERT-based models are pre-trained on large-scale textual data and are time- and energy-consuming to extract features from documents in the target corpus.

3.2. Parameter setting and topic extraction

Due to the effectiveness and efficiency of doc2vec when dealing with topic extraction tasks in Section 3.1, it was applied for topic extraction and further scientific evolution analysis. There are two models in the doc2vec architecture. One is the Distributed Bag-of-Words (DBOW) model and the other is Distributed Memory of paragraph vectors (DMPV). Although model selection for doc2vec has long been discussed, no agreement has been reached on the performance difference between DBOW and DMPV. Lau and Baldwin [54] found that DBOW, despite being the simpler model, not only requires fewer training epochs but also performs better; while Le and Mikolov [17] reported that DMPV is superior. In accordance with the former benchmarks, we selected the DBOW model to perform doc2vec.

Existing studies on embedding models and their applications have pointed out that hyperparameter optimization can be as important as, or even more important than, the choice of models themselves [55]. Most scientometric studies choose hyperparameters according to experience, testing, and best practices reported in the literature, since no objective consensus has been reached on optimizing hyperparameters [22, 56]. Curiskis, et al. [23] pointed out that the performance of doc2vec varied significantly for the number of training epochs, which is subject to the document lengths. The smaller the document sizes, the more epochs are needed. Mendsaikhan, et al. [57] observed that window size settings, as well as subsampling rate, only have a slight effect on document similarity tests, but the tuning epochs or the vector size influences the effectiveness of doc2vec. We draw on the experience of the benchmarks conducted above, considering that all these documents are abstracts of scientific literature, which have similar document sizes and syntax features. We mapped all documents into 300-dimension vectors, with the training epochs set to 10 and the window size set to 5.

All documents were represented as vectors after doc2vec training, which were then classified into t piles according to their publication time. For the publications of each year, the k-means method was applied. Documents with similar semantics were then grouped together, which can be regarded as a topic. One of the most important concerns when applying k-means to cluster scientific textual data lies in the setting of the K value. For the first year's documents, K was determined by measuring the Sum of Squared Errors (SSE). And for the subsequent $t-1$ years, we referenced the research conducted by Fortunato, et al. [58], assuming that the number of topics covered by articles is proportionate to the number of unique terms contained in their titles. Therefore the K values were determined by counting unique title terms in each year's documents.

3.3. Scientific evolutionary paths unraveling

3.3.1. Topic designation and "Head-Body" identification

After topic extraction, for each topic, the most representative terms were extracted for topic designation, following the processes shown in Fig. 2. Inspired by embedding-based topic network

construction and measurement [59], we employed word2vec to create topological structures of words for each topic since it performs well in capturing semantics in Section 3.1, and used network-based metrics to extract the most important terms to designate topics.

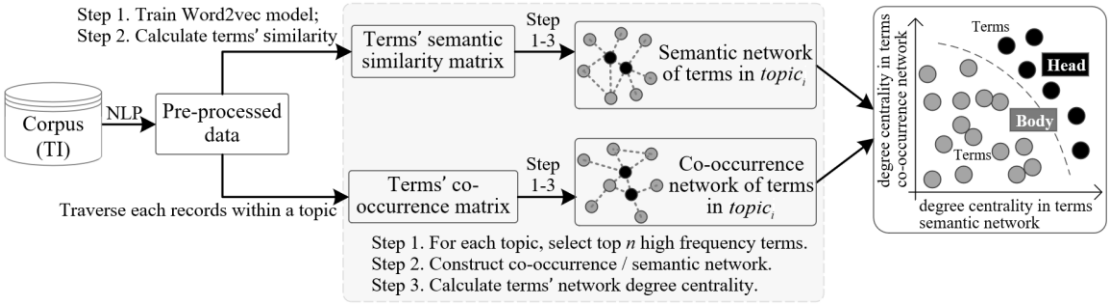

Fig. 2 Process of selecting the most representative terms for topic designation

Note: Terms with higher network degree centrality (m are picked from the semantic network and the other m are from the co-occurrence network) are painted as black dots, here we call them the ***Head*** of a topic. And its' ***Body*** is formed by all terms within this topic.

Titles of documents were first collected as a corpus. We cleaned this corpus to eliminate punctuations, non-alphabetic characters, stop words, and commonly used academic words. The remaining terms were then lemmatized and consolidated, and those that occur more than two times were retained [60]. We input the cleaned data into the word2vec training module. There are two models for word2vec: Skip-Gram and Continuous Bag-of-Words (CBOW), which show similar performance according to benchmarks conducted by Levy, et al. [55]. We selected the CBOW model to convert each term into a δ-dimensional numeric vector.

The top n high-frequency terms within each topic were then used to build the semantic network and co-occurrence network of terms. In the semantic network, the weight of edges between every two nodes indicates the semantic similarity of corresponding terms, which is determined based on the cosine similarity of terms' vectors extracted using word2vec. We pruned the semantic networks to remove weak relationships. And for terms' co-occurrence network, the weight of network edges is equal to the co-occurrence frequency of every two terms.

Using the Python NetworkX toolkit[††], we calculated the degree centrality indicator for terms in semantic and co-occurrence networks, respectively, to gauge the "importance" of terms in each topic [3]. As shown in Fig. 1, terms with higher degree centrality in either of the two networks are highlighted in black, here we call them the *Head* of a topic. It can be used for manual topic designation. And its *Body* is formed by all terms within this topic.

3.3.2. Evolutionary states identification for linked topics

When mapping the scientific evolutionary paths with extracted topics, the Pearson's Correlation Coefficient (PCC) was used to calculate similarities between topic vectors – or say the body of topics – across different periods [61, 62]. The PCC matrix of topic vectors in every two connected years is calculated as follows:

$$PCC_{(T_{t-1,i},T_{t,j})} = \frac{\sum_{d=1}^{\gamma}\left(T_{t-1,i,d} - \overline{T_{t-1,i}}\right)\left(T_{t,j,d} - \overline{T_{t,j}}\right)}{\sqrt{\sum_{d=1}^{\gamma}\left(T_{t-1,i,d} - \overline{T_{t-1,i}}\right)^2}\sqrt{\sum_{d=1}^{\gamma}\left(T_{t,j,d} - \overline{T_{t,j}}\right)^2}} \quad i,j = 0,1,2,\ldots$$

[††] https://networkx.org/

where $T_{t-1,i}$ represents the topic vector i in the year T_{t-1}, and $T_{t,j}$ represents the topic vector j in the year T_t. These vectors are computed as the average of document vectors since each topic is a cluster of semantically related documents. As shown in Fig. 3, totally $t-1$ PCC matrices are acquired for t years' topics. To highlight strong linkages and identify evolutionary states, only correlation coefficients that larger than the upper quartile are kept [59].

Considering both the head consistency and the body similarity of topics in vector spaces across different periods, four evolutionary states are identified, with a sample and some description provided in Fig. 3 and Table 3:

1) Newborn: a topic is newborn if its head has never appeared before.
2) Knowledge convergence: this means that there might be diffusion, interaction, and integration between the predecessor topic and its successor. The two topics share similar bodies, but have unique heads. For example, there is a great content overlap between topics of 'clustering' and 'classification' – they may share some theories and concepts – they are two different topics with knowledge convergence. Quantitatively, if the similarity of two topic bodies is greater than 90%, meanwhile they have different topic heads, we define that they have knowledge convergence links.
3) Knowledge inheritance: this means that the predecessor and the successor topic are the same ones, which not only have high semantic similarity in their bodies but also share the same head. Quantitatively, if the similarity of two topic bodies is greater than 90%, meanwhile, they have the same topic head, we define them as having knowledge inheritance links.
4) Extinction: a topic is dead if its head no longer appears.

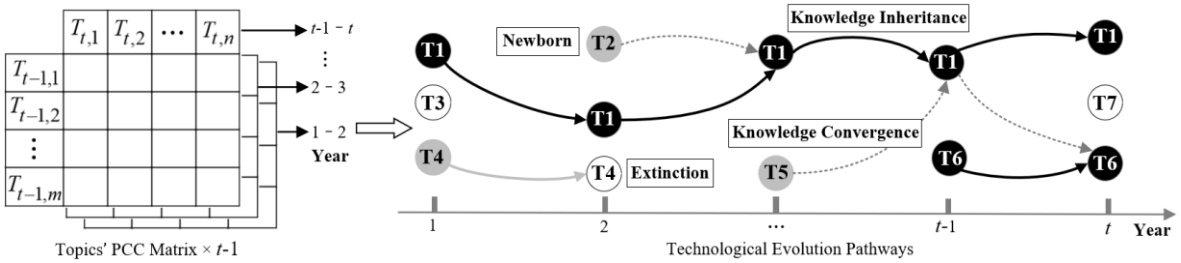

Fig. 3 Adjacency matrices constrcution and scientific evolution mapping

Table 3 Four types of topics' evolutionary states

Evolutionary states	Description
Newborn	Topics without precursors
Knowledge convergence	The precursor and successor have different heads but similar bodies
Knowledge inheritance	The precursor and successor share the same heads and similar bodies
Extinction	Topics without successors

4. Case study: scientific evolution analysis on object detection research

4.1. Data

To demonstrate the feasibility and effectiveness of the methodology in unraveling scientific

evolutionary paths, we select one of the fastest evolving fields in artificial intelligence research – object detection – as the target area to conduct an empirical case study. Object detection aims to provide semantic understanding of images and videos, and has been widely studied to support face recognition, autonomous driving, human behavior analysis, etc. [63]. As a long-standing and challenging research area in computer vision, it has attracted considerable attention over the past decade.

The dataset used in this study is collected from the Web of Science database, with a strategy of searching title, abstract, and keyword fields for terms including "object detection," "edge detection," "saliency detection," "face detection," "pedestrian detection," "change detection," "feature detection," "anomaly detection," "corner detection," and "motion detection." A collection of 56,529 peer-reviewed documents published between 2011 and 2020 was retrieved. A detailed search strategy is provided in Appendix C.

As shown in Fig. 4, the number of object detection papers has climbed sharply. However, the expansion of research topics, which are determined by counting unique terms of article titles, maintains a steadier pace.

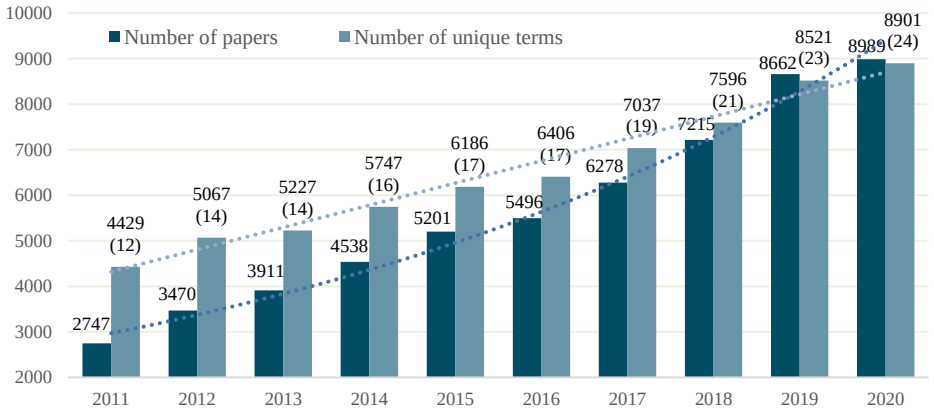

Fig. 4 Annual production of scientific papers in WoS, and the number of unique terms (with the number of topics noted in parentheses) covered by these articles.

4.2. Scientific evolutionary path identification

We applied doc2vec and k-means algorithms to extract topics from object detection publications. The number of topics in 2011 is set as 12, which increases to 24 by the year 2020. Topics' heads are used for depicting themselves, according to the method we have proposed in section 3.3.1. Several topic designation examples are shown in Fig. 5. As we can observe, the head of face detection contains terms such as face detection, facial, face recognition, Haar, and AdaBoost. In particular, both the Haar classifier and the AdaBoost algorithm are widely used in face detection tasks [64]. The head of network traffic anomaly detection consists of anomaly detection, network, network traffic, traffic anomaly, and intrusion detection. This research theme mainly focuses on the problem of detecting anomalies and attacks in network traffic. As for landscape change detection, besides change detection, terms including Landsat, land use/cover, and vegetation are also highlighted. Details about topics are provided in Table 10, Appendix D.

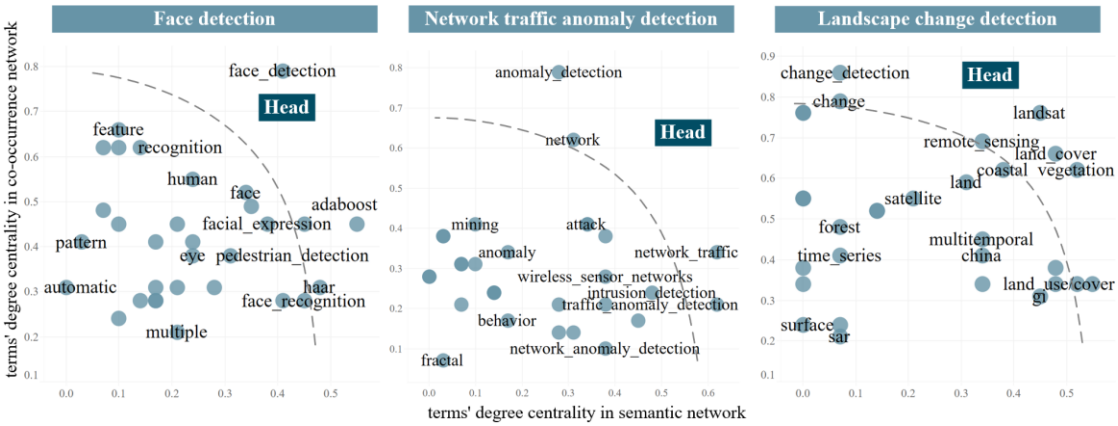

Fig. 5 Examples of topic designation

Fig. 6 maps the topic evolutionary paths of object detection research. The dotted lines represent knowledge integration, while the solid lines imply knowledge inheritance. If the precursor and the successor topics are connected with a solid line, they are the same topics. In total, 39 topics appeared between 2011 and 2020, involving a series of important algorithms, hardware implementations, and practical applications.

As we can observe, object detection technology has powered many aspects of modern society, such as security, medicine, transportation, environment, and military, as early as 2011. Over the past decade, sustained attention has been paid to *face detection* (2011), *medical image segmentation* (2011), *networks traffic anomaly detection* (2011), *landscape change detection* (2011), and so forth. Some classical detection methods, including *adaptive edge detection* (2011), *saliency detection* (2011), and *corner detection* (2011), have also attracted considerable research interest[‡‡].

More recently, with the explosive growth of multisource, multimodal, and unstructured data, deep convolutional neural networks have brought about a revolution in object detection [65]. The development of *FPGA* (2015), which is configurable, flexible, and energy-efficient, meets the high computational demands of *deep CNNs* (2016) [66]. Some cutting-edge models and frameworks appear, such as *You Only Look Once (YOLO)* (2019) and *Single Shot MultiBox Detector (SSD)* (2018), offering superior accuracy and real-time efficiency for real-word object detection applications [63]. Fig. 6 highlights some prominent achievements that arise along with massive data and deep learning techniques, including *data-driven anomaly detection* (2018), *deep learning-based pedestrian detection* (2017), *anomaly detection in IoT* (2017), and so on. In addition, advanced object detection techniques also assist the development of *autonomous driving* (2019) and *smart navigation* (2020). Impressively, scientists have also begun to emphasize the importance of *weakly supervised object detection* (2019). Since CNN-based detectors rely heavily on large-scale annotated data, it is essential to explore how to leverage the power of CNNs when only weakly annotated data are accessible [63].

‡‡ The time when topics debuted is given in parentheses.

Fig. 6 Scientific evolutionary paths in the field of object detection

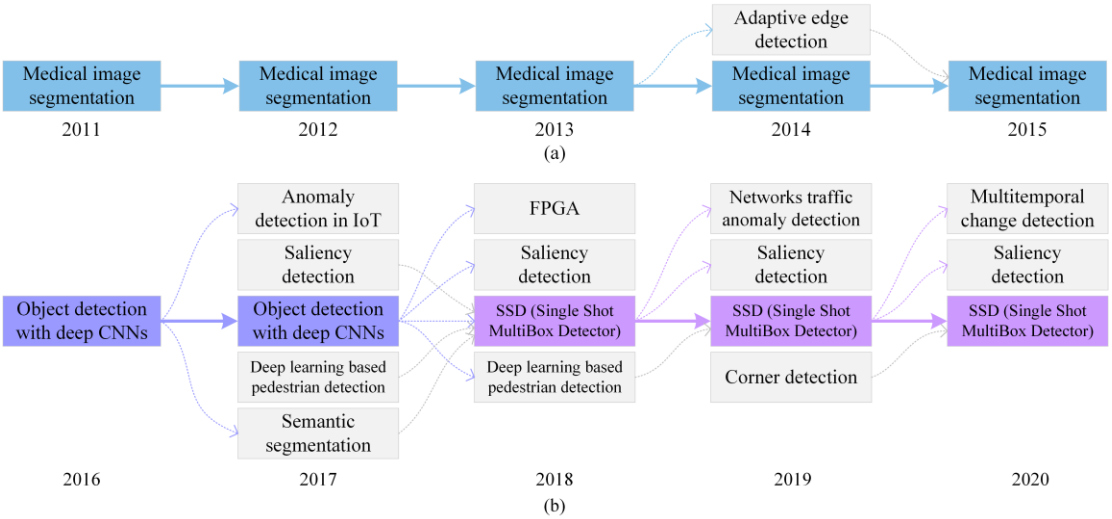

Fig. 7 Examples of four evolutionary states. (a) mainly knowledge inheritance, (b) mainly knowledge integration, with a topic newborn in 2018 (*SSD*) and a topic extinction in 2017 (*Object detection with deep CNNs*).

The scientific evolutionary paths depict the characteristics and patterns of knowledge diffusion, interaction, and convergence within the field of object detection. New ideas come in, and old topics die out. To better understand evolutionary states among topics, we choose three topics as examples and illustrate their process of knowledge flow. As shown in Fig. 7(a), *medical image segmentation* is a relatively independent research topic, which has contributed greatly to medical applications with technologies such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [67]. However, topics such as *object detection with deep CNNs* and *Single Shot MultiBox Detector (SSD)* are more likely to absorb or disseminate existing knowledge. As demonstrated in Fig. 7(b), *deep CNNs* are dead in 2017, however, it does not mean that CNNs have gone out of the scientists' focus – *SSD* is emerging in the next year. *SSD* combines ideas from Faster R-CNN and multiscale CONV features, achieving both high detection quality and fast detection speed [63]. This kind of evolution can be regarded as a process of knowledge recombination and technological advancement.

4.3. Expert assessment and evaluation

To evaluate how meaningful the results obtained are and to provide a further interpretation of the topic changes over time, we invited 20 domain experts from both academia and industry to quantitatively evaluate the above results. A questionnaire was designed to validate the topic evolution result in the field of object detection, which involves two parts: the quality assessment of embedding-based topic extraction and the accuracy of the subsequent evolutionary paths identification.

The first part focuses on evaluating the quality of embedding-based topic extraction to identify new topics within a continuous period. We record the debut year for all the extracted topics to label their novelty. Domain experts actively involved in computer vision research were then invited to mark the novelty of randomly selected topics and give scores from 1 to 5, where 1 stands for completely mature, 2 means comparatively mature, 3 represents neutral, 4 means comparatively novel, and 5 stands for novel. An average novelty score (ANS) can be calculated for each topic to

quantify expert cognition. As we can observe in Table 4, the topics that have emerged in recent years have relatively high ANS. However, for a few topics that denote specific algorithms or models, such as SSD (Single Shot MultiBox Detector) and YOLO (You Only Look Once), fewer agreements have been reached between experts and the proposed method. One possible reason is that scientists in a fast-growing and developing research field like object detection are faced with a high-speed update of algorithms and approaches. Therefore, approaches that were proposed several years ago are mature and even outdated for them.

The second part of the questionnaire is to validate whether all embedding-based evolutionary paths fit expert cognitions. For a number of randomly selected topic pairs, experts were invited to assess their relevancy and mark from 1 to 3. Here, 1 means that the two topics are weakly correlated, 2 stands for neutral, and 3 represents strongly correlated. An average relevance score (ARS) was calculated for each pair of topics on the pathways to quantify expert opinion. As shown in Table 5, the pair of topics identified as related in the evolutionary paths have relatively high ARS. It was shown that embedding-based evolutionary paths identification, in general, has the capability of capturing topic relations in different periods.

Table 4 The average novelty scores and the debut years of topics.

Topics	Debut year	Average novelty score	Pass evaluation
Face detection	2011	1.30	YES
Data driven anomaly detection	2018	2.90	YES
Corner detection	2011	2.15	YES
Weakly supervised object detection	2019	2.85	YES
Medical image segmentation	2011	2.10	YES
Smart navigation	2020	3.30	YES
Single Shot MultiBox Detector (SSD)	2018	2.65	NO
You Only Look Once (YOLO)	2019	2.00	NO

Table 5 The average relevance score and the identified relevance of each pair of topics.

Topic pairs	Connected in evolutionary path	Average relevance score	Pass evaluation
Auditory change detection – medical image segmentation	NO	1.20	YES
Object detection with deep CNNs – SSD (Single Shot MultiBox Detector)	YES	2.80	YES
Anomaly detection in multivariate data – network traffic anomaly detection	YES	2.80	YES
3d point cloud – climate change	NO	2.05	YES
Face detection – pedestrian detection	YES	2.40	YES

Wearable sensor – FPGA (Field Programmable Gate Array)	NO	2.00	YES
Adaptive edge detection – saliency detection	YES	2.45	YES

5. Conclusions and future research

Uncovering scientific evolutionary paths can help scientists stay abreast of ever-changing research trends and hotspots. Furthermore, it can provide funding agencies with insights into the transformation of the scientific paradigm, enabling them to enhance funding strategies accordingly. The fundamental aspects of identifying scientific evolution analysis from a text mining perspective encompass topic extraction, connection analysis, and identification of evolutionary states. Although previous research has achieved fruitful outcomes through topic extraction using bibliometric methods or bag-of-words-based document representation models, important limitations remain in effectively distinguishing evolutionary states.

The primary contribution of this paper is the proposal of a novel embedding-based methodology for scientific evolution analysis. This methodology proves effective in extracting topics, measuring semantic similarities between topics, and quantitatively distinguishing the evolutionary states of topics. To ensure the superiority of word and document embedding in topic extraction and scientific evolution analysis, we first benchmarked the performance of four candidate embedding methods with BOW-based models. The results of the study indicate that doc2vec outperforms other methods and demonstrates its effectiveness and efficiency in capturing semantics from textual data. While SciBERT does a good job when dealing with scientific publications, it is time- and energy-consuming, therefore, might be an effective substitute for doc2vec when users have access to GPUs. Based on the performance of these embedding methods, doc2vec and word2vec are applied for topic extraction and scientific evolution analysis, with the assistance of NLP techniques, a clustering algorithm, and a network analysis indicator. With full consideration of the consistency and similarity of topics in vector spaces, four topic evolutionary states – newborn, knowledge convergence, knowledge inheritance, and extinction – are distinguished using the "Head-Body" method we proposed, making it possible to quantitatively measure and understand the knowledge flow of topics, or branches of science, in the form of changes in topic vectors over time.

The present study has certain limitations that warrant further investigation. Firstly, we utilized k-means clustering to group document vectors into topics based on its established superiority in previous research. Future research might validate whether it is preferable from an empirical perspective. Moreover, we did not incorporate dimensionality reduction techniques prior to clustering, which has the potential to improve the performance of vector clustering. This aspect will be considered in future research to further enhance the study's findings. In addition, we did not take into account the inner network structures of the topics. Network analysis indicators, such as density, average path length, and clustering coefficient, may potentially provide additional insights into the identification of the evolutionary states of topics and are worth more attention and effort in future research.

References

[1] Y. Qian, Y. Liu, and Q. Z. Sheng, "Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence," *Journal of Informetrics,* vol. 14, no. 3, p. 101047, 2020/08/01/ 2020.

[2] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *Journal of Informetrics,* vol. 14, no. 3, 2020.

[3] A. Zeng *et al.*, "The science of science: From the perspective of complex systems," (in english), *Physics Reports,* vol. 714, pp. 1-73, 2017.

[4] S. Jung and W. C. Yoon, "An alternative topic model based on Common Interest Authors for topic evolution analysis," *Journal of Informetrics,* vol. 14, no. 3, 2020.

[5] H. Y. Xu, J. Winnink, Z. H. Yue, Z. Q. Liu, and G. T. Yuan, "Topic-linked innovation paths in science and technology," *Journal of Informetrics,* vol. 14, no. 2, MAY 2020.

[6] H. Liu, Z. Chen, J. Tang, Y. Zhou, and S. Liu, "Mapping the technology evolution path: a novel model for dynamic topic detection and tracking," *Scientometrics,* vol. 125, no. 3, pp. 2043-2090, 2020/12/01 2020.

[7] M. Katsurai and S. Ono, "TrendNets: mapping emerging research trends from dynamic co-word networks via sparse representation," *Scientometrics,* vol. 121, no. 3, pp. 1583-1598, DEC 2019.

[8] M. Mariani and M. Borghi, "Industry 4.0: A bibliometric review of its managerial intellectual structure and potential evolution in the service industries," *Technological Forecasting and Social Change,* vol. 149, DEC 2019.

[9] J. Gläser, W. Glänzel, and A. Scharnhorst, "Same data—different results? Towards a comparative approach to the identification of thematic structures in science," *Scientometrics,* vol. 111, no. 2, pp. 981-998, 2017/05/01 2017.

[10] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics," *Journal of the Association for Information Science and Technology,* https://doi.org/10.1002/asi.23814 vol. 68, no. 8, pp. 1925-1939, 2017/08/01 2017.

[11] H. S. Chen, G. Q. Zhang, D. H. Zhu, and J. Lu, "Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014," (in english), *Technological Forecasting and Social Change,* vol. 119, pp. 39-52, Jun 2017.

[12] L. Aristodemou and F. Tietze, "The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data," *World Patent Information,* vol. 55, pp. 37-51, DEC 2018.

[13] Z. L. Yang, N. Islam, Y. N. Shi, K. Venkatachalam, and L. C. Huang, "The Evolution of Interindustry Technology Linkage Topics and Its Analysis Framework in Three-Dimensional Printing Technology," *IEEE Transactions on Engineering Management,* 2021.

[14] J. Ramos, "Using TF-IDF to determine word relevance in document queries," presented at the Proceedings of the first instructional conference on machine learning, 01/01, 2003.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," (in english), *Journal of Machine Learning Research,* Article; Proceedings Paper vol. 3, no. 4-5, pp. 993-1022, May 2003.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems,* pp. 3111-3119, 2013.

[17] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," presented at the Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume

32, Beijing, China, 2014.

[18] H. S. Chen, Q. Q. Jin, X. M. Wang, and F. Xiong, "Profiling academic-industrial collaborations in bibliometric-enhanced topic networks: A case study on digitalization research," *Technological Forecasting and Social Change,* vol. 175, Feb 2022, Art. no. 121402.

[19] L. Aristodemou, "Identifying Valuable Patents: A Deep Learning Approach," Department of Engineering, University of Cambridge, 2021.

[20] S. Azimi, H. Veisi, M. Fateh-rad, and R. Rahmani, "Discovering Associations Among Technologies Using Neural Networks for Tech-Mining," (in English), *IEEE Transactions on Engineering Management,* vol. 69, no. 4, pp. 1394-1404, AUG 2022.

[21] M. R. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint*, vol. abs/2203.05794, Available: https://arxiv.org/abs/2203.05794

[22] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," *Neurocomputing,* vol. 266, pp. 336-352, Nov 2017.

[23] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Information Processing & Management,* vol. 57, no. 2, p. 102034, 2020/03/01/ 2020.

[24] H. Miao, Y. Wang, X. Li, and F. F. Wu, "Integrating Technology-Relationship-Technology Semantic Analysis and Technology Roadmapping Method: A Case of Elderly Smart Wear Technology," *IEEE Transactions on Engineering Management,* vol. 69, no. 1, pp. 262-278, FEB 2022.

[25] K. Hu *et al.*, "Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis," *Information Processing & Management,* vol. 56, no. 4, pp. 1185-1203, 2019/07/01/ 2019.

[26] X. Li, Q. Xie, T. Daim, and L. Huang, "Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology," *Technological Forecasting and Social Change,* vol. 146, pp. 432-449, 2019/09/01/ 2019.

[27] E. Garfield, I. H. Sher, and R. J. Torpie, "The Use of Citation Data in Writing the History of Science," 1964.

[28] N. P. Hummon and P. Doreian, "Connectivity in a citation network: The development of DNA theory," (in English), *Social Networks,* vol. 11, no. 1, pp. 39-63, MAR 1989.

[29] Y. Huang, F. J. Zhu, A. L. Porter, Y. Zhang, D. H. Zhu, and Y. Guo, "Exploring Technology Evolution Pathways to Facilitate Technology Management: From a Technology Life Cycle Perspective," *IEEE Transactions on Engineering Management,* vol. 68, no. 5, pp. 1347-1359, OCT 2021.

[30] L. Chen, S. Xu, L. J. Zhu, J. Zhang, H. Y. Xu, and G. C. Yang, "A semantic main path analysis method to identify multiple developmental trajectories," (in English), *Journal of Informetrics,* vol. 16, no. 2, MAY 2022, Art. no. 101281.

[31] H. Park and C. L. Magee, "Tracing Technological Development Trajectories: A Genetic Knowledge Persistence-Based Main Path Approach," *PLoS One,* vol. 12, no. 1, p. e0170895, 2017.

[32] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences,* vol. 99, no. 12, pp. 7821-7826, 2002/06/11 2002.

[33] S. Fortunato, "Community detection in graphs," *Physics Reports,* vol. 486, no. 3, pp. 75-174, 2010/02/01/ 2010.

[34] W. Pan, L. Jian, and T. J. S. Liu, "Grey system theory trends from 1991 to 2018: a bibliometric analysis and visualization," *Scientometrics,* vol. 121, pp. 1407 - 1434, 2019.

[35] I. Majdouline, J. E. Baz, and F. Jebli, "Revisiting technological entrepreneurship research: An updated bibliometric analysis of the state of art," *Technological Forecasting and Social Change,* vol. 179, p. 121589, 2022/06/01/ 2022.

[36] Y. Zhang, G. Q. Zhang, H. S. Chen, A. L. Porter, D. H. Zhu, and J. Lu, "Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research," (in english), *Technological Forecasting and Social Change,* vol. 105, pp. 179-191, Apr 2016.

[37] B. Song and Y. Suh, "Identifying convergence fields and technologies for industrial safety: LDA-based network analysis," (in english), *Technological Forecasting and Social Change,* vol. 138, pp. 115-126, Jan 2019.

[38] C. E. Moody, "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec," *arXiv: Computation and Language,* Available: https://arxiv.org/abs/1605.02019

[39] Y. Zhai, Y. Ding, and H. Zhang, "Innovation adoption: Broadcasting versus virality," *Journal of the Association for Information Science and Technology,* https://doi.org/10.1002/asi.24420 vol. 72, no. 4, pp. 403-416, 2021/04/01 2021.

[40] I. Yamashita, A. Murakami, S. Cairns, and F. Galindo-Rueda, "Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundsat initiative," 2021.

[41] Y. Zhang *et al.*, "Does deep learning help topic extraction? A kernel k-means clustering method with word embedding," (in english), *Journal of Informetrics,* vol. 12, no. 4, pp. 1099-1117, Nov 2018.

[42] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *EMNLP/IJCNLP*, 2019: Association for Computational Linguistics.

[43] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2019.

[44] Y. Wang, Y. Hou, W. Che, and T. Liu, "From static to dynamic word representations: a survey," *International Journal of Machine Learning and Cybernetics,* vol. 11, no. 7, pp. 1611-1630, 2020/07/01 2020.

[45] S. Woo, J. Youtie, I. Ott, and F. Scheu, "Understanding the long-term emergence of autonomous vehicles technologies," *Technological Forecasting and Social Change,* vol. 170, p. 120852, 2021/09/01/ 2021.

[46] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint,* Available: https://arxiv.org/abs/2008.09470

[47] M. J. Sanchez-Franco and M. Rey-Moreno, "Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings," *Psychology & Marketing,* vol. 39, no. 2, pp. 441-459, FEB 2022.

[48] H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," (in English), *Education and Information Technologies,* 2023.

[49] X. G. Wang, J. He, H. Huang, and H. Y. Wang, "MatrixSim: A new method for detecting the evolution paths of research topics," *Journal of Informetrics,* vol. 16, no. 4, NOV 2022, Art. no. 101343.

[50] H. Wu, H. F. Yi, and C. Li, "An integrated approach for detecting and quantifying the topic evolutions of patent technology: a case study on graphene field," *Scientometrics,* 2021.

[51] H. M. Zhu, L. Qian, W. Qin, J. Wei, and C. Shen, "Evolution analysis of online topics based on 'word-topic' coupling network," (in English), *Scientometrics,* vol. 127, no. 7, pp. 3767-3792, JUL 2022.

[52] Q. Wang and L. Waltman, "Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus," *Journal of Informetrics,* vol. 10, no. 2, pp. 347-364, 2016/05/01/ 2016.

[53] A. M. Dai, C. Olah, and Q. V. Le, "Document Embedding with Paragraph Vectors," *arXiv preprint,* Available:

https://ui.adsabs.harvard.edu/abs/2015arXiv150707998D

[54] J. H. Lau and T. Baldwin, "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation," presented at the Association for Computational Linguistics, 2016. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv160705368L

[55] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the association for computational linguistics,* vol. 3, pp. 211-225, 2015.

[56] B. Thijs, "Using neural-network based paragraph embeddings for the calculation of within and between document similarities," (in English), *Scientometrics,* Article vol. 125, no. 2, pp. 835-849, Nov 2020.

[57] O. Mendsaikhan, H. Hasegawa, Y. Yamaguchi, and H. Shimada, "Identification of Cybersecurity Specific Content Using the Doc2Vec Language Model," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 2019, vol. 1, pp. 396-401.

[58] S. Fortunato *et al.*, "Science of science," *Science,* vol. 359, no. 6379, p. eaao0185, 2018/03/02 2018.

[59] H. Chen, X. Song, Q. Jin, and X. Wang, "Network dynamics in university-industry collaboration: a collaboration-knowledge dual-layer network perspective," *Scientometrics,* 2022/03/19 2022.

[60] Q. Jin, H. Chen, X. Wang, T. Ma, and F. Xiong, "Exploring funding patterns with word embedding-enhanced organization–topic networks: a case study on big data," *Scientometrics,* vol. 127, no. 9, pp. 5415-5440, 2022/09/01 2022.

[61] B. N. Yan, T. S. Lee, and T. P. Lee, "Mapping the intellectual structure of the Internet of Things (IoT) field (2000-2014): a co-word analysis," *Scientometrics,* vol. 105, no. 2, pp. 1285-1300, NOV 2015.

[62] P. Jafarzadeh and F. Ensan, "A semantic approach to post-retrieval query performance prediction," *Information Processing & Management,* vol. 59, no. 1, JAN 2022, Art. no. 102746.

[63] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 261-318, 2020/02/01 2020.

[64] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, 2004, pp. 79-84: IEEE.

[65] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015.

[66] S. Mittal, "A survey of FPGA-based accelerators for convolutional neural networks," *Neural Computing and Applications,* vol. 32, no. 4, pp. 1109-1139, 2020/02/01 2020.

[67] A. Norouzi *et al.*, "Medical Image Segmentation Methods, Algorithms, and Applications," *Iete Technical Review,* vol. 31, no. 3, pp. 199-213, MAY-JUN 2014.

[68] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," *Information Sciences,* vol. 477, pp. 15-29, 2019/03/01/ 2019.

[69] A. Vaswani *et al.*, "Attention is all you need," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint,* Available: https://arxiv.org/abs/1810.04805

[71] G. A. Ronda-Pupo and J. S. Katz, "The scaling relationship between citation-based performance and coauthorship patterns in natural sciences," *Journal of the Association for Information Science and Technology,* https://doi.org/10.1002/asi.23759 vol. 68, no. 5, pp. 1257-1265, 2017/05/01 2017.

Appendix

Appendix A: Introduction of word2vec, doc2vec, and BERT

The input of word2vec is a sequence of words: $X = \{w_{i-k}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+k}\}$, in which w_i stands for a target word and k is the context size of it; the window size is $2k + 1$. This paper selects the CBOW model to train word2vec. According to the method proposed by Mikolov, et al. [16] and the notation given by Kim, et al. [68], the main objective of word2vec is to maximize the average log probability $L(D)$ when predicting the target word, in which the probability $P(w_i|w_{i-k}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+k})$ is formulated with a SoftMax function:

$$L(D) = \frac{1}{\varphi}\sum_{i=1}^{\varphi} \log P(w_i|w_{i-k}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+k})$$

$$P(w_i|w_{i-k}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+k}) = \frac{e^{y_{w_i}}}{\sum_j e^{y_j}}$$

in which φ is the size of corpus, indicating the number of unique terms, y_j is the j^{th} output value of a feed-forward neural network computed using the equation below:

$$y = b + Uh(w_{i-k}, \dots, w_{i+k}; W)$$

where b, U denote the bias terms and weight matrix between the hidden and output layers; h is constructed by averaging words vectors, which are extracted from the word embedding matrix W.

Doc2vec inherits the same main idea as word2vec. At prediction time, document vectors are generated by fixing the word vectors and training the new vector until convergence [17].

BERT is a transformer-based model pre-trained with MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) objectives on a large corpus. It expects input data in a specific format – each sentence is supposed to start with a special token [CLS] and end with [SEP]. The core of BERT is the adoption of the transformer technique [69]. According to the model architecture developed by Devlin, et al. [70], transformer relies on a multi-head attention mechanism, where several attention layers run in parallel. The input vectors, consisting of query (Q), keys (K), and values (V), are transferred to Scaled Dot-Product Attention through linear projection, which can be formulated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Appendix B: Benchmark datasets, indicators, and model parameters

The datasets used for document clustering tasks are supposed to have a predefined taxonomy, with topics, subject categories, or disciplines as labels. Therefore, we collected documents from two bibliometric datasets, the one is Web of Science (WoS) and the other is MedLine, covering both multidisciplinary and mono-disciplinary classification systems. The composition of two datasets are shown as tables 6 and 7.

Table 6 WoS dataset

Research fields	Web of Science category (WC)
Arts and humanities	Philosophy; History
Social science	Economics; Education & Educational Research
Life science and biomedicine	Immunology; Plant Sciences

Nature science	Mathematics; Optics
Applied science	Computer Science, Artificial Intelligence; Mechanics

Table 7 MedLine dataset

Root MeSH	Leaf MeSH
Anatomy	Intestine, Large
Organisms	Bacteriophages
Diseases	Corneal Diseases
Chemicals & Drugs	Dental Materials
Psychiatry & Psychology	Depressive Disorder
Biological Sciences	Biocatalysis
Physical Sciences	Nanomedicine
Technology & Food & Beverages	Dairy Products
Humanities	Philosophy
Information Science	Telecommunications

Three metrics were used to compare the clustering results of document vectors generated by different embedding methods. The evaluation measures should be independent of the absolute values of the labels; in other words, a permutation of the cluster label values will not change the score value in any way [23]. Therefore, metrics such as Rand Index, Mutual Information, and Fowlkes-Mallows can be appropriate.

The Rand Index (RI) evaluates clustering performance by considering the proportion of right predicted samples. The mathematical formulation of RI is illustrated in the following way, where $a + d$ is equal to the number of agreeing pairs. To be more specific, a $(or\ d)$ is the count of pairs that fall within the same (or different) clusters in ground truth labels and predicted labels. Both b and c indicate the number of disagreeing pairs. b is the count of pairs that are not grouped in the same cluster but belong to the same category according to the truth labels, while c represents the count of pairs that are assigned in the same cluster but belong to different categories.

$$RI = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}} = \frac{2(a+d)}{n(n-1)} \tag{5}$$

The Fowlkes-Mallows (FM) score is algebraically equivalent to the geometric mean between precision and recall. According to the definitions above, FM can be measured as follows:

$$FM = \frac{a}{\sqrt{(a+c)(a+b)}} \tag{6}$$

Mutual information (MI) quantifies the dependence between two random variables. The following formula shows the calculation process of MI, where $p(x)$ and $p(y)$ are the marginal probability density functions of variables X and Y, while $p(x,y)$ represents their joint probability density function. MI equals one when two random variables are completely independent. When it comes to a clustering result, perfect labeling is scored as one, and it gets zero if all labeling is false.

$$MI(X,Y) = \sum_{y\in Y}\sum_{x\in X} log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{7}$$

Considering that the MI doesn't guarantee that random label assignments will get a value close to zero, the Adjusted Mutual Information (AMI) is then proposed:

$$AMI(X,Y) = \frac{MI(X,Y) - E[MI(X,Y)]}{\max[H(X),H(Y)] - E[MI(X,Y)]} \tag{8}$$

where $H(X)$ and $H(Y)$ are marginal entropy, and $E[MI(X,Y)]$ denotes the expected value of MI. Similarly, the Adjusted Rand Index is given by:

$$ARI(X,Y) = \frac{RI(X,Y) - E[RI(X,Y)]}{max[RI(X,Y)] - E[RI(X,Y)]} \tag{9}$$

Both the ranges of AMI and ARI are between -1 and 1. The larger the value, the better the clustering effect.

In this paper, we use FM, AMI, and ARI to measure the clustering effect, and then compare the performance of different embedding methods.

To minimize the impact of hyperparameters on overall performance, we test all models with different sets of hyperparameter settings. The finally used parameter settings are listed in Table 8.

Table 8 Parameter setting for different models. Default values in the gensim toolkit are noted in parentheses.

Model	Parameter setting
LDA	iterations = 6000 (50), chunksize = 5000 (2000), passes = 40 (1), alpha = 'auto', eta = 'auto'
Word2vec	vector size = 300 (100), epochs = 100 for WoS dataset and 50 for MedLine (5), min count = 5
Doc2vec	vector size = 300 (100), min count = 5, epochs = 10 (5)

Appendix C: Search Strategy of object detection papers

The collection of "object detection" papers is a multistep process as summarized in Table 8. We first retrieve publications within the scope of six well recognized journals and conferences in the field of computer vision, including CVPR, ICCV, ECCV, IJCV, TIP, and TPAMI. This step leads to a total of 37288 records. Keywords of these articles are listed in descending order by frequency, we then pick out the top 10 keywords which contain the term "detection" to formulate the second search query. We follow Ronda-Pupo and Katz [71] to retrieve all peer-reviewed documents, including article, note, proceedings papers, and review. Finally, a collection of 56529 documents published between 2011 and 2020 are inclusive.

Table 9 The multistep data collection process

Database: Web of Science core collection database	
Editions: Science Citation Index Expanded (SCIE); Conference Proceedings Citation Index-Science (CPCI-S)	
STEP 1	Search Query: (SO = ("IEEE Transactions on Pattern Analysis and Machine Intelligence" or "IEEE Transactions on Image Processing" or "International Journal of Computer Vision") OR CF = ("International Conference on Computer Vision and Pattern Recognition" or "International Conference on Computer Vision" or "European Conference on Computer Vision" OR "ICCV"

	OR "ECCV" OR "CVPR")) AND LA = (English)[8] Publication Date: No time limitation Date of Export: 2021-11-15 **Number of Records: 37288** (17631 journal papers and 19657 conference papers)
STEP 2	Search Query: TS = ("object detection" or "edge detection" or "saliency detection" or "face detection" or "pedestrian detection" or "change detection" or "feature detection" or "anomaly detection" or "corner detection" or "motion detection") AND LA = (English) AND DT = (Article OR Note OR Review OR Proceedings Paper) Publication Date: Between 2011-01-01 and 2020-12-31 Date of Export: 2021-11-16 **Number of Records: 56529**

1

[8] This search query should be refined by removing conferences such as "International Conference on Computer Vision and Graphics," "International Conference on Computer Vision Theory and Applications," and so on.

Appendix D: Topics extracted from object detection articles

Table 10 Topics extracted from object detection articles

Global ID	Topics	Content and local ID
1	adaptive edge detection	edge detection, edge, adaptive | noisy image, smoothing, edge detection P1-01, P2-12, P3-11, P4-13, P5-06, P6-13, P7-16, P8-19, P9-13, P10-21
2	face detection	feature, face detection, real time | AdaBoost, cascade, face recognition P1-02, P2-14, P3-10, P4-08, P5-03, P6-02, P7-04, P8-16, P9-16, P10-08
3	radar	sensor, CMOS, motion detection | CMOS, array, chip P1-03, P2-04, P3-03, P4-02, P5-13, P6-17, P8-13, P9-08, P10-02
4	real time detection	mobile, real time, video | multi camera, camera, robot P1-04, P2-05, P6-07, P7-15, P8-09, P9-19
5	saliency detection	object detection, segmentation, shape | appearance, discriminative, saliency P1-05, P2-08, P3-04, P4-09, P5-08, P6-04, P7-06, P8-10, P9-12, P10-14
6	auditory change detection	processing, motion, auditory | auditory, stimulus, cortex P1-06, P2-06, P3-07, P4-16, P5-07, P6-11, P7-05, P8-17, P9-18, P10-05
7	medical image segmentation	automatic, motion, 3d | MRI, optical coherence, cardiac P1-07, P2-03, P3-05, P4-03, P5-15, P6-05, P7-13, P8-15, P9-14, P10-06
8	networks traffic anomaly detection	network, anomaly detection, clustering | network traffic, wireless sensor networks, intrusion detection P1-08, P2-09, P3-08, P4-15, P5-09, P6-09, P7-18, P8-05, P9-20, P10-24
9	multitemporal change detection	change detection, automatic, building | multitemporal, hyperspectral images, hyperspectral P1-09, P2-13, P3-02, P4-10, P5-01, P6-08, P7-14, P8-02, P9-04, P10-09
10	corner detection	edge detection, wavelet, automatic | canny, edge detection, defect P1-11, P2-10, P3-06, P4-01, P5-10, P6-12, P7-10, P8-18, P9-01, P10-19
11	landscape change detection	change, change detection, monitoring | landscape, vegetation, wetland P1-12, P2-01, P3-14, P4-12, P5-02, P6-10, P7-08, P8-20, P9-23, P10-01

12	video detection	tracking, camera, video | moving object, moving object detection, video surveillance P1-10
13	moving object detection	video, local, tracking | foreground, background subtraction, moving object P2-11, P3-01, P4-06, P5-11, P6-03, P7-17, P8-21, P9-03, P10-22
14	real time DSP	real time, video, implementation | embedded, DSP, GPU P2-02
15	sequential anomaly detection	anomaly detection, change detection, network| sequential, spectrum sensing, Bayesian P2-07
16	human detection	recognition, human, video | robot, mobile robot, tracking P3-12, P7-01
17	accelerator	implementation, real time, architecture | processor, accelerator, embedded P3-09
18	sequential change detection	change detection, anomaly detection, network | sequential, spectrum sensing, wireless sensor networks P3-13, P4-14
19	real time processing	real time, architecture, fast | low power, processor, FPGA P4-07, P5-05
20	3d point cloud	real time, 3d, object detection | point cloud, stereo, camera P4-05, P5-04, P6-15, P7-12, P8-11
21	pedestrian detection	object detection, pedestrian detection, object | deformable part, occlusion, cascade P4-04, P5-16
22	laser Doppler vibrometer	radar, sensor, measurement | Doppler, vibration, beam P4-11
23	anomaly detection in multivariate data	anomaly detection, network, change detection | quickest change detection, multivariate, online P5-12, P6-01, P7-07, P8-03, P9-02, P10-23
24	FPGA	architecture, embedded, FPGA | accelerator, FPGA, embedded P5-17, P6-16, P7-19, P8-12, P9-07, P10-12

25	video anomaly detection	video, anomaly detection, network | service, knowledge, safety P5-14
26	wearable sensor	sensor, strain sensor, stretchable | highly sensitive, stretchable, graphene P6-14, P7-02, P8-06, P9-22, P10-11
27	object detection with deep CNNs	object detection, CNN, fast | object proposal, multi view, proposal P6-06, P7-11
28	anomaly detection in IoT	anomaly detection, network, real time | IoT, distributed, service P7-09, P8-08, P9-21, P10-20
29	deep learning based pedestrian detection	object detection, classification, deep learning | car, pedestrian, camera P7-03, P8-07, P10-16
30	data driven anomaly detection	anomaly detection, network ,graph | anomalous, social network, social P8-14, P9-15, P10-07
31	wind turbine condition monitoring	anomaly detection, monitoring, network | forecasting, condition monitoring, wind turbine P8-04, P9-05, P10-04
32	autonomous driving	object detection, 3d, object | slam, autonomous driving, monocular P9-10, P10-15
33	weakly supervised object detection	object detection, deep, network | weakly supervised, shot, semi supervised P9-09, P10-10
34	YOLO	deep learning, object detection, CNN | augmentation, YOLOv3, object detector P9-17, P10-13
35	climate monitoring	extreme, basin, China | precipitation, north, climate change P9-06
36	smart navigation	recognition, object detection, smart | indoor, real time, robot P10-17
37	hyperspectral anomaly detection	anomaly detection, low rank, hyperspectral anomaly detection | low rank, hyperspectral anomaly detection, sparse representation

		P10-18
38	SSD (Single Shot MultiBox Detector)	r CNN, net, single shot | object detection, network, deep P8-01, P9-11, P10-03
39	semantic segmentation	semantic, RGB, attention | human, object, object detection P7-01