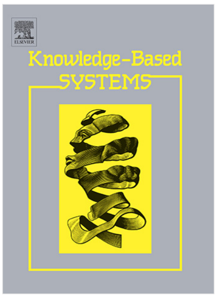

Identifying a common pattern within ancestors of emerging topics for pan-domain topic emergence prediction

Sukhwan Jung, Aviv Segev

PII: S0950-7051(22)01113-3
DOI: https://doi.org/10.1016/j.knosys.2022.110020
Reference: KNOSYS 110020

To appear in: *Knowledge-Based Systems*

Received date : 1 April 2022
Revised date : 11 October 2022
Accepted date : 12 October 2022

Please cite this article as: S. Jung and A. Segev, Identifying a common pattern within ancestors of emerging topics for pan-domain topic emergence prediction, *Knowledge-Based Systems* (2022), doi: https://doi.org/10.1016/j.knosys.2022.110020.

Identifying a Common Pattern within Ancestors of Emerging Topics for Pan-Domain Topic Emergence Prediction

Sukhwan Jung[a,1], Aviv Segev[a]

[a]*Department of Computer Science, University of South Alabama, 150 Student Services Dr, Mobile, 36688, AL, USA*

Abstract

The shared interest among existing research topics matures over time until it emerges as a topic of its own. This paper detects emerging topics as well as general predictor models spanning multiple research domains through the network-based topic evolution approach, which offers additional topic evolution capabilities such as extrapolation of data and separation of topic transition and correlation. Topics are represented as their neighbors in the past, or ancestors, and their structural properties are used to train binary classification models in capturing the materialization of such topics. The entirety of 197 million publications within the Microsoft Academic Graph was used to build multiple datasets, where machine learning algorithms were trained with structural features resulting in over 0.98 area under the precision-recall curve. General topic emergence predictor equations are then proposed based on the models trained specifically for each domain, which were able to capture a common pattern shared by emerging topics in general.

Keywords: Topic Prediction, Scientometrics, Knowledge management, Machine learning

1. Introduction

Research is a collective work, where scientists expand the currently available knowledge by contributing discoveries in the form of publications. The gradual expansion based on past knowledge is a foundation of valid and sound research activities. Knowing the boundary of knowledge is therefore essential for any researcher to provide meaningful contributions; one needs to know

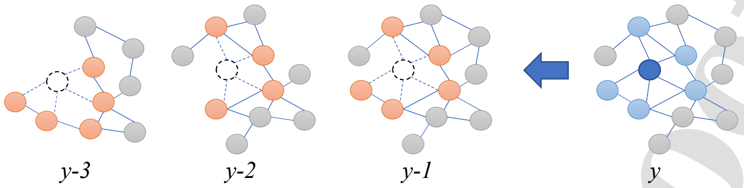

Figure 1: Linking an emerging topic (blue) in year y with their ancestors (red) in previous years.

where the boundary is before expanding on it. Such information has both academic and industrial value by providing insight on setting the research goal; research efforts can be directed towards the more valuable discoveries, or instead be invested in the future technologies yet to be studied by the competitors. A summarized overview of research topics would alleviate the researchers' workload, as it helps with internalizing current knowledge in the research domain. Topic evolution automatically detects and tracks research topics by employing text-based topic models over document collections.

The network-based topic emergence detection is illustrated in Figure 1, where neighbors of an emerging topic in year y are tracked in previous years through their neighbors or *ancestors*. The authors have previously shown that ancestors of emerging topics are structurally distinguishable from ancestors of already existing topics [1]. This paper aims to validate the assumption that emerging research topics over different domains share a common graphical pattern in their ancestors, and a single classifier model can be built to identify and predict emerging topics, in other words, topics added to the given network, regardless of the searched domains. Such a model needs to be semantic-independent to work on different domains, therefore a network-based topic evolution approach is employed where topics are defined by their neighbors in a topic network. This allows context assignment to new topics in the future where related documents are yet to be published, as the past relationships between their ancestors can be utilized instead. Relationships between different topics and semantic transitions within a single topic are also separated, as the edge and node attributes respectively reflect such changes. This allows the identification and prediction of more evolutionary event types such as merge and split on the evolving networks, which are hard to distinguish using text-based topic models. The proposed approach is first applied to the list of top-level research domains on a publicly available bibliographic dataset, utilizing multiple machine learning techniques to capture the possibly distinct patterns in different datasets. The trained models are then utilized to propose a generalized topic emergence prediction model which can universally be applied to any research domain.

The shared interest among existing topics matures over time until it materializes as a topic of its own; the maturation and ultimate materialization of such topics are reflected on the topic network as a new node with the contributing topics as its ancestors. The proposed method incorporates this behavior and classifies topic subgraphs as whether a new topic will emerge from them, with relationships within the subgraph representing interactions leading up to the new topic. Various research domains are experimented on to cover different interests and scopes, ranging from specific techniques or phenomena to research fields such as art, science, and engineering methods. The top two levels in the hierarchical fields-of-study ontology from the Microsoft Academic Graph dataset were used to represent a total of 311 research domains within 197.64 million publications, containing 18.67 million overlaps from a set of 0.71 million topics.

The problem is formalized as classifying new nodes with the graphical features of their ancestors. Multiple machine learning classifiers are trained to capture various domain-specific research behaviors; for example, the research activities in *philosophies* are different from those of *quantum physics*. The experiment result showed that binary classification is capable of identifying ancestors of emerging topics. Logistic regression models performed the best in more than 1/3 of the datasets, indicating the data are linear in many cases. The models trained for each domain are then analyzed to extract a domain-independent general pattern as a form of a single regression model. The assumption is that emerging topics' ancestors share a set of common structural features throughout different domains. The authors generated a pan-domain topic emergence predictor regression model, showing that the general predictor model can successfully predict topic emergence in a majority of the domains using only four out of the total 15 features used during training. The exceptions were several medical and economics-related domains where the common model failed, indicating that there are domains with topic co-occurrence patterns distinct from the rest of the research communities.

Section 2 reviews the related work on topic evolution, previous attempts at the prediction of new topics, as well as background research for the proposed method. Sections 3 and 4 detail the proposed method and experimentation, and the experiment results are shown in Section 5. Finally, Section 6 lists the concluding remarks and future works.

2. Related Work

Topic evolution is the field of research focusing on the temporal evolution of topics, where topics represent the shared theme within the given set of text documents. *Content transitions* of a single topic and *topical correlations* between different topics are analyzed over time [2] into one of the six event types including *survive*, *dissolve*, *grow*, *shrink*, *split*, and *merge* [3]. The traditional approaches mainly define topics as probabilistic models such as Latent Dirichlet Allocation and its variants. Topics are first independently extracted from temporally ordered subcollections to form a series of timeslot-specific topic models [4]. The topic models on consecutive timeslots are then connected with similarity measures, where differences between connected topics are viewed as evolution within the connected topics. Dynamic topic models [5] used a fixed timeslot approach to capture the evolution of chained topics. Evolutionary theme pattern mining [6] expanded on this approach by allowing multiple connections between sequentially ordered topics over time instead of looking at only one-to-one connections. A chain of similar topics over consecutive timeslots is then interpreted as the existence of content transition between the similar topics [7]. Another recent approach involves the two-tier topic model method, where the global topic models spanning across the whole timeslots are used as temporal anchors between multiple local topics over time [8]. The static global topics are used as the topic evolution branches on which multiple local topics can join using cosine similarities with a threshold. Evolutionary transition is found by changes in sizes and numbers of local topics connected to a single global topic over the years. This however had a limitation of measuring only evolutionary transitions within time-insensitive global topics.

Topic models represent topics by a series of word vectors, which are used to represent the word identity and topical context at the same time [9]. Evolution within a single topic and evolution involving multiple topics are therefore indistinguishable as they are all represented by the same high semantic similarities. The emergence of topics is generally not considered in the traditional topic evolution because of the semantic-based similarity measures as well, as topics are never truly semantically *unique* in the distribution-based topic models. Manual interpretation is required to detect such evolution. There are numerous attempts at topic evolution with the use of non-textual data to overcome such a limitation. The bibliographical relationships between publications and their authors have been proposed to enhance topic evolution.

The evolutionary transition was augmented with cross-citation count within topic membership documents [10]. Using a citation context was proposed to deal with the small number of documents assigned to topics as well, expanding the document collection with documents cited by its members [11]. The citation information was also used to overcome the topic dilution problem, utilizing cited document semantics to distinguish autonomous parts with originalities from inherited contents [12]. This resulted in topic models with more emphasis on the research front of the publications. The research front consists of new terms with sharp frequency growth, or emerging topics.

Applications of topic evolution with different topic definitions have also been studied in recent years. New topic identification [13] defined topics as the users' interests. The query intervals and semantic patterns during the search engine querying session are used to identify topics, where new topics are found with distinct search patterns. Online recommender systems aim to identify and predict changes in user interests based on previous purchases and browsing records. The recent approach in recommender systems is the use of knowledge graphs [14], where graph completion and embedding techniques are utilized to provide more accurate recommendations with better explainability. Technology forecasting [15] aims to predict future technology characteristics. Communities in a temporal keyword co-occurrence network replaced the distribution-based topic models [16], using the pre-defined medical subject headings dataset from PubMed as keywords. Keyword clusters are considered topics, and similarities between temporally neighboring clusters were measured to identify the evolutionary transitions.

These researches are focused on alleviating the fundamental limitation of the text-based topic models, where the identity and content of a topic are represented by the same feature. The two-tiered topic model method was tested on 700 thousand publications over 20 years using LDA topic models as it is one of the recent approaches focusing on merge and split events [17]. Text-based topic evolution was able to capture distinct major topics in the dataset but performed poorly at detecting merge and split events between them. The majority of topics with focused interests fail to appear in any meaningful topic evolution paths as the evolution tracks were dissolved into time-spanning global topics; topics evolve over 20 years with enough changes that static global topics cannot be semantically connected to each of them. Few semantically generic topics dominated the evolution paths as they were most often used together as background materials; three out of ten global topics were responsible for more than 90% of the merge and split events.

Network-based topic evolution approach is proposed to separate the identity and content of a topic, allowing access to more accurate topic merge and split detection. The author-based topic evolution method was proposed where the multiple author relationships within the bibliographic dataset are used to capture fluid evolution between changing topics [17]. Multigraph clustering is applied to five-dimensional author networks, and each author group is defined as the representation of the topic they have a common interest in. Topic evolution identification based on the topic co-occurrence network is proposed to retain the versatility of network-based topic evolution while removing the uncertainty and external factors of the author relationships [18]. The emergence of new topics was identified by defining the topic as its neighbors in the topic network, where the previous relationships between the neighbors were classified with machine learning techniques. Merge and split events were also captured, showing continuous evolution paths between topics over time as well as semantic transitions made within a given topic at the same time.

3. Network-Based Topic Emergence Prediction

Statistical topic models rely on the word distributions in the given document collection. This approach is capable of identifying already present topics in the research field but is limited to retrospective identification. Prospective evolutionary prediction, such as emerging topic detection, requires the knowledge of the novel topic models which need to be extracted from the yet non-existent future documents. This limitation is further exacerbated by the fact that a topic's identity is directly tied to its contents, which are used to semantically connect topics over time. Continued evolution within a single topic and evolution involving multiple topics are both measured by semantic similarities, rendering them hard to be distinguished. The proposed method utilizes network structures instead to overcome such limitations, where an identity of a topic and relationships between topics are separated as nodes and edges in topic co-occurrence networks. Such separation allows the method to simultaneously capture *content transition* and *topical correlation*, detecting complex topic evolution events such as merge, split [17], and emergence. The extrapolative nature of evolving networks also grants the possibility of prospective prediction for events in the future, without accessing relative data from the target timeslot. Topic emergence is a complex topic evolution event that is often impossible to predict using text-based methods on non-existent publications; this paper aims to show that the proposed method can capture such events as well. Emerging topics are represented as new nodes, where the context of new topics can be derived from the contexts and previous behaviors of their neighboring nodes.

3.1. Topic Emergence Identification

Research topics cannot be accurately represented without considering the research domains they are observed in, as the topics' statuses are dependent on when and where they are used. For example, a fully matured research topic in one domain could be transferred into another domain as an aspiring topic. The popularity of a research topic can vary in different domains as they have unique research backgrounds and interests. The emergence of new topics is therefore measured within the scope of specific research domains, each represented by a specific research topic called *domain topic*. Topic network $G_{d,y}$ is generated for each domain topic d by extracting topics and topical co-occurrences found from a d-related document collection in a given year y. Each topic v present in the vertex set $V_{d,y} = (v)$ has at least one co-occurrence with d, and the edge set $E_{d,y}$ represents the topic co-occurrence frequencies between vertex v_1 and v_2 at y with frequency $w_{d,y}$ as the weight.

Ancestors of an emerging topic are represented as a group of topics with a previously unseen predecessor in the network. Neighbors are first found for each topic in $G_{d,y+1}$ to generate a set of nodes connected to the target future topic. They are then projected to $G_{d,y}$ to extract a set of nodes $V'_{d,y}(v)$ which will have a single topic v as their common predecessor in the next year. Any ancestors that are not present in the given year's topic network are disregarded. The ancestors $V'_{d,y}(v)$ are then categorized based on whether the predecessor is *new* or *old* in $y + 1$. For a given domain topic d, there is a fixed set of topics $v \in V_d$ used over the years. Each topic v has a specific year when it was first used within the domain, and $init_d(v)$ is defined to represent the year when v first co-occurred with d. The given topic v is *new* when $init_d(v) = y + 1$. Topics previously appearing in earlier topic networks are defined as *old* when $init_d(v) < y + 1$. The binary state of a given topic v at year y in domain topic d is defined as $c_{d,y}(v)$ which is calculated as the inverted ceiling function applied on the normalized differences between $init_d(v)$ and y. The *new* topics are denoted as $c_{d,y}(v) = 1$, while *old* topics are denoted as $c_{d,y}(v) = 0$. $V'_{d,y}(v)$ is built for each node set in $V_{d,y+1}$, which can reach over 100,000 depending on the size of the research domain. Filtering is done to reduce the number of analyzed subgraphs. As subgraph sizes follow the power law, the long tail as well as the short head are filtered out as they are either too minor to be meaningful or too broad for meaningful analysis. Firstly, extreme size differences in the few largest $V'_{d,y}(v)$ were mediated by filtering *new* and *old* topic groups larger than the maximum size in the other group. Then the top n largest topic groups $t_{d,y}(v)$ were selected for each label to form a maximum of $2n$ topic groups.

$$\begin{aligned} V'_{d,y,n} = \{&max_n(new) \cup max_n(old) \mid new, old \in V'_{d,y}(v), v \in V_{d,y+1}, \\ &|new| \leq max(|V'_{d,y}(v)|) \text{ where } c_{d,y}(v) = 0, \\ &|old| \leq max(|V'_{d,y}(u)|) \text{ where } c_{d,y}(u) = 1\} \end{aligned} \tag{1}$$

Different research domains exhibit varying research patterns over time a single ML algorithm might not be able to capture. Several algorithms were therefore deployed for binary classifications within a given domain topic d with varying combinations of the number of topic groups to classify (n), and the length of training years (l). For a given year y, classifier models are trained using topic groups for l previous years $\{T_{d,i,n} \mid y - l \leq i < y\}$, using top n rows for each label. A multi-layered ANN method using stochastic gradient descent was used to train an adaptive deep learning (DL) [19] model with a cross-entropy loss function and rectifier activation function. A DL model generates in-model features and hence the interpretation of the result is harder, but it can deal with less optimized feature sets. Distributed Random Forest (DRF) generates multiple random forest learners trained on partial data and uses the average outcome for the final prediction [20]. While it is generally considered a good predictor where the data imbalance between labels is inevitable, DRF also suffers from low interpretability due to the complex nature of the process. A gradient boosting machine (GBM) uses parallel tree boosting and iteratively builds models for each branch. The models are sequentially trained based on the outcome of their predecessors to minimize the cross-entropy loss function. The XGBoost framework [21] (XGB) is a more regularized version of GBM for computational resource efficiency and lessening the over-fitting problem, which has many hyperparameters to control and hence is harder to tune albeit with better performances and outlier control. Lastly, logistic regression (LR) provides a high degree of result interpretability without hyperparameter tuning despite its linear capability.

Structural features of the topic groups are measured to be used in the binary topic emergence classification. 15 subgraph-related, node-related, edge-related, and weighted features are selected as dependent variables. They represent different dimensions of information stored within networks, which are later pruned with feature engineering techniques.

3.2. Identifying General Patterns of Emerging Topics

The classification models are trained for top n topics for each domain topic d at year y, generating models specifically trained for the given topic groups $T_{d,y,n}$. Any model trained by one set of topic groups $T_{d,y,n}$ can be applied to other topic groups with different domain topics, years, or group sizes $T_{d',y',n'}$ as all of the models share the same data structure with the identical set of features; no modifications to the model are necessary. Successful binary classifications for $T_{d',y',n'}$ would indicate that the binary labels in $T_{d,y,n}$ and $T_{d',y',n'}$ share a common pattern, which was captured when the model was trained with one set of data. Utilizing the trained model with other testing data would provide the generalizability of the method over different datasets.

The averaged linear predictor equation is then proposed. A logistic regression model can be described as a single formula $e^{x^\top\beta+\beta_0}/1+e^{x^\top\beta+\beta_0}$ with the input feature data x, the feature coefficient set β, and the intercept β_0. Averaging β and β_0 over multiple models provides a general model that is common across different topic groups. The general model will represent the common emerging topic pattern between different domain topics instead of showing the global background pattern that is seen in general research.

4. Experiments

4.1. Dataset

Microsoft Academic Graph (MAG) [22] is selected from many publicly available bibliographic datasets from which to extract topic networks. The MAG is one of the more recently created bibliographic datasets but is competitive with other major datasets such as Google Scholar and Scopus datasets in size and coverage [23]. The MAG also provides a hierarchical concept ontology called Fields-of-Study (FoS) [24]. The six-level concept ontology renews monthly by applying a series of graph link analyses and convolutional neural networks to Wikipedia articles. The concepts are then tagged to the publication weekly with the help of large-scale multilevel text classification. This is important in this research as assigning topics from dataset-wide topic vocabularies is a large task in itself. The FoS are already extracted within the MAG dataset for the indexed publications and therefore are used as topics assigned to the publications in the experiment.

Table 1: Summary of the 311 Datasets for Each Domain Topic.

	#FoS	#Papers	#PaperFoSlinks
min	1,264	553	4,770
mean	60,024	1,351,113	9,405,866
max	484,664	24,546,680	169,418,047
std	50,420	2,755,881	19,397,927

The February 2020 snapshot of the MAG dataset is downloaded through Microsoft Azure Databricks, including 197,642,464 publications, 709,934 FoS, 48,829 journals, more than 1.5 billion citation links, and over 1.3 billion paper-FoS links. The preprocessing is done to extract partial datasets, as the topic network over the whole domain would be too complex to retrieve and process. Bibliographic records related to a specific FoS are extracted to represent a subset collection, sharing the same topic. The purpose of this paper is to show that there is a common topic emergence pattern spanning various research domains, therefore all FoS with the top two highest levels in the hierarchical concept ontology are used. The outcome is a total of 311 sub-datasets covering all 197 billion publications, each focused on one of the 19 level 0 FoS or 292 level 1 FoS. Some level 1 FoS such as *arthistory* and *organicchemistry* represent subsets of the relative level 0 FoS (*art* and *chemistry*) where an almost complete hierarchy can be observed. There are also level 1 FoS with less hierarchical roots such as *botany* or *law*, representing the smaller yet fairly independent research fields.

Table 1 shows the average size of FoS-specific datasets. A total of 18.67 million FoS are used with all 311 topic networks, each network on average having 8.45% of the total FoS. The topic networks generally represent interdisciplinary research fields connected to a wide range of topics, sharing many common topics between them. Commonalities between topic networks are also reflected in the overlaps ratio, as topics on average appeared over 26.3 domains while only 2.13 overlaps were observed per publication. There are also large variances in the number of topics and publications related to them, with the largest FoS dataset $d = medicine$ alone consisting of 324,171 topics from 24.55 million papers, and nearly 169 million links between them. This is a stark difference from the smallest FoS dataset $d = ceramic\ materials$ having only 4,770 links. This allows the experiment to be run on a wide range of datasets, which have different sizes and structural features.

4.2. Topic Emergence Identification

The dataset preprocessing is done using the Alabama Supercomputer Authority[1] high performance computing service, converting the totality of the MAG dataset into 311 separate datasets. Topic groups $T_{d,y,n}$ using the Ancestor topics in (1) are then extracted from the network. Various domain topics have different histories, and therefore the datasets at the same y do not represent the domains with the same research maturities. It was still deemed more appropriate to compare topic groups in the same year as the overall changes in the research behaviors with the technological development and expansion of research communities outweigh the length of history in the field. Ancestors of each topic v in the next year $y + 1$ are extracted, and the size-ranked subgraphs in y are filtered by n for the *new* and *old* topics.

The experiment involves running multiple classifications using thousands of training set combinations. It is impractical to optimize or engineer all possible ML models; therefore an automated ML framework is used for the classification tasks. H2O AutoML[2] is one of the leading open-source Automated Machine Learning (AutoML) interfaces, providing automated access to a variety of basic and complex ML algorithms. The AutoML is set up to train ten models for each ML task with ten-fold cross validation, using DL, DRF, GBM, XGB, and LR as the possible candidates and previous l topic networks as training sets. It then combines the resulting trained models to build two stacked ensemble (SE) models [25]. SE_{best} is computed by voting the best-performing models from each ML algorithm family, while SE_{all} utilizes all ten models. The features have different value ranges and therefore were standardized to have zero mean and unit variance.

Pilot experiments from journal-specific datasets [18] were referenced to filter out hyperparameters y, n, and l to remove a large number of unnecessary iterations over 311 domain topics. Consecutive years showed minimal changes to the trained models; hence $y = [2000, 2005, 2010, 2015]$ were used instead, excluding the incomplete 2020. $n = [100, 250, 500]$ and $l = [1, 5, 9]$ were used as they showed significant performance differences within specific ranges $0 \leq l \leq 10$ and $100 \leq n \leq 500$. A total of 11,196 $T_{d,y,n,l}$ were generated as a result, training a total of 134,352 classification models with ten individual ML models and two stacked ensemble models.

[1] https://hpcdocs.asc.edu/

[2] https://h2o.ai/platform/h2o-automl/

4.3. Identifying General Patterns of Emerging Topics

The LR model has shown high performance in the experiment, and the general patterns were identified by building the averaged linear predictor function $lp = x^T u\beta' + \beta'_0$ using the trained models. The coefficients $\beta' = \frac{1}{m} \cdot \beta J_{m,1}$ and the intercept $\beta'_0 = \frac{1}{m} \cdot \beta_0 J_{m,1}$ are averaged over m models by multiplying the matrices with the vector of ones J. β and β_0 from the trained LR models cannot be directly applied to the raw data as they are the result of the standardized feature set. The standardization is reversed to retrieve *normal* coefficients and intercept, de-scaling the value and subtracting the added offsets. These are not identical to the coefficients from the model trained with no standardization, which often experiences overfitting issues with features with larger scales. The logistic regression function is then rounded to generate the binary classification result; 0 for $False$, and 1 for $True$. The datasets are extended to encompass 20 years in the 21st century $y = [2000, \ldots, 2019]$ with less number of algorithms used, resulting in a total of 44,784 additional models.

5. Results

5.1. Topic Emergence Identification

The topic emergence identification results were measured for different $T_{d,y,n}$ to analyze the performance over a different combination of variables. The model with the lowest *logloss* was selected out of twelve ML models trained for each training set given, and average values for 311 d were shown in Table 2. The summarized result shows that the proposed method is able to identify topic groups associated with new future topics with high accuracy ranging around 0.99 *auc* and *aucpr*, even using only topic groups in the directly previous year as the training set. *Logloss*, *mse*, and *rmse* all show a distinctive decrease with $l > 1$, nearly halving the values. This indicates that the likelihood of the correct prediction increases with a larger training set size. The differences made by l are less pronounced and are statistically insignificant with $p > 0.1$ for all measures. This indicates that the method works with varying dataset sizes, retaining almost all of its classification power with 1/5 of the largest dataset used.

While different T show high performances, not all ML models are successfully trained. Four out of the five lowest performing domain topics

Table 2: Performances of the Best-performing Topic Emergence Identification Models.

l	n	*auc*	*aucpr*	*logloss*	*mpce*	*mse*	*rmse*
1	100	0.9900	0.9902	0.2426	0.0221	0.0543	0.1905
	250	0.9897	0.9897	0.2773	0.0237	0.0625	0.1976
	500	0.9896	0.9882	0.2778	0.0240	0.0634	0.1984
5	100	0.9910	0.9912	0.1172	0.0205	0.0288	0.1474
	250	0.9906	0.9906	0.1120	0.0225	0.0290	0.1496
	500	0.9904	0.9891	0.1110	0.0232	0.0290	0.1503
9	100	0.9909	0.9911	0.1098	0.0208	0.0284	0.1477
	250	0.9904	0.9903	0.1087	0.0227	0.0283	0.1493
	500	0.9902	0.9887	0.1057	0.0233	0.0280	0.1499

d=['ceramic materials', 'classical economics', 'Keynesian economics', 'polymer science'] have the fewest topic groups, indicating that the limited number of topic groups can result in incorrect classifications. *Ceramic materials* is showing the least successful classification results with over 0.4 *mpce* and 0.54 *auc* score; this is because this topic is not fully indexed on the MAG snapshot used in the experiment. As the smallest dataset used, d = *ceramic materials* results in zero topic groups since $y = 2017$ as no publication is linked to it. Even the non-empty subsets in $y < 2017$ have on average 70.1 topics per year when up to 1,000 are retrieved for others. This is not the reflection of actual research communities as around 18,800 related publications since 2017 were indexed in Google Scholar, and there are nine related publications in the same time slot on the up-to-date live MAG dataset query result. It is clearly an outlier and therefore is removed from further analysis.

There are domain topics that result in a very low result on specific training sets. For *d=operation_management*, LR resulted in $auc(T_{2010,250}) = 0.0997$ and $auc(T_{2010,500}) = 0.3164$ while the average *auc* of other T reached 0.9894. Similar issues can be found with other algorithms, as well. When all trained models are considered, $auc(T_{materials\_science,2015,250}) = 0.3418$ and $auc(T_{geology,2015,500}) = 0.4655$ can be seen for DRF with $l = 1$, as well as $auc(T_{business,2015,500}) = 0.5638$ for GBM with $l = 5$. The repeated runs result in a trained model with $auc > 0.9$ with identical algorithms for all such instances, showing that the labels may be too clearly separated causing regression models not to work properly. The stark differences shown within

Table 3: Percentages of Classification Algorithms which Resulted in the Lowest *logloss*.

l	n	DL	DRF	GBM	LR	SE	XGB
	100	9.52%	4.03%	14.35%	38.39%	22.42%	11.29%
1	250		2.34%	14.44%	45.48%	28.95%	8.79%
	500		1.21%	11.37%	46.69%	31.13%	9.60%
	100		3.39%	23.06%	33.71%	18.55%	21.29%
5	250		1.94%	15.97%	46.13%	18.39%	17.58%
	500		0.65%	15.40%	50.65%	20.73%	12.58%
	100		3.47%	22.74%	35.48%	17.58%	20.73%
9	250		2.10%	19.27%	41.85%	17.74%	19.03%
	500		0.81%	18.39%	47.58%	18.55%	14.68%

the same domain can also be attributed to the limitation of the proposed method; it cannot effectively deal with topic groups with multiple common predecessors. Topic groups are identified solely based on the neighborhoods of future topics; therefore identical topic groups will be generated for both labels if a new topic and an existing topic share the same ancestors. The classification will fail when there are multiple cases of duplicate topic groups, as it's trying to find differences between identical data.

The performance of ML algorithms showed no significant differences when the top-rank models were compared. One-way Anova for the models with rank 1 for each of the l, n combinations resulted in statistically insignificant mean differences with $p = 0.25$ and $p = 0.43$ for *auc* and *aucpr* respectively. As there are no major differences between the algorithms, one is selected to be used for further analysis. Table 3 shows the percentage of ML algorithms selected as best for each of the topic emergence classification processes; there were insignificant differences between SE_{all} and SE_{best} in terms of performance and therefore they are merged into one column. The table shows that 33.7% of $T_{d,y,n,l}$ at minimum had the best results with LR models. The dominance of LR over the SE indicates that the topic group data are heavily linear, as the ensemble model is considered to be the strong model for classifying complex data. It is also worth noting that DL and DRF were very rarely ranked first. This is because they are in the more complex spectrum of binary classification and their full potential cannot be reached with unengineered models. The limited extrapolation capabilities with the unseen

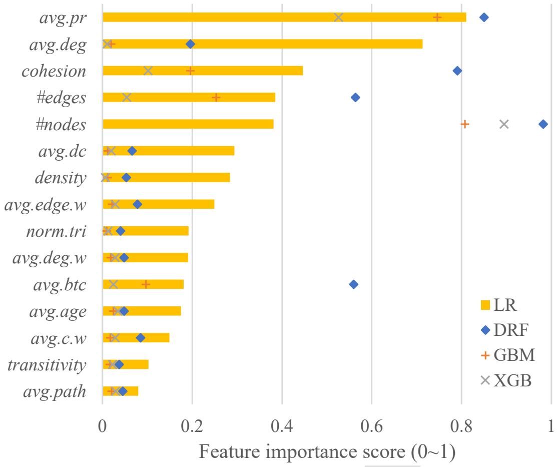

Figure 2: Averaged feature importance trained from $T_{d,y,500,9}$.

data can also be the cause of the low performance. The topic group features have large variances, implying that there will be several test data with the unobserved feature ranges during the training process.

Feature importance analysis revealed that LR also applies distinctive importance to the 15 features used, shown by case studies using $n = 500$ and $t = 9$ which is a combination with the least *logloss* as shown in Table 2. Figure 2 shows the difference between ML algorithms in terms of feature importance using $T_{d,y,500,9}$, representing LR as bars and the rest as dots. DL is not shown in the figure as it generates a set of internal features instead; SE utilizes DL models and therefore it is also omitted. The figure shows there are distinctive differences between LR and other algorithms. *#nodes* is heavily weighted in the other algorithms but shows only 42.6% of the averaged importance in LR. This indicates that the topic emergence prediction is not dependent on the topic neighbor sizes, which is often considered one of the major quality metrics for clusters. LR shows much higher interest for other features instead, putting 9.43, 9.1, and 11.65 times the weight to *avg.deg*, *avg.dc*, and *density* respectively compared to the rest. Similar patterns can be found in all $T_{d,y,n,l}$. LR is shown to have the best results in general and also has distinct feature importance patterns showing its unique role in the given classification tasks. LR was selected to identify general patterns of emerging topics in the next section; hence it is used for further analysis.

The performance of the LR models stayed relatively high when not ranked at the top with the average *aucpr* score of 0.9819. The average *logloss = 0.2971* is comparably worse compared to the performances of the best-performing models as shown in Table 2, but this is reduced down to 0.1197 when more topic networks are used for training with *l = 9*. Excluding *ceramic materials* and a few other outliers, performance metrics over 311 domain topics with up to 1,000 topic groups over four year periods resulted in high *auc* values with low *mpce* and *mse* indicating that both the true positives and true negatives were successfully captured.

5.2. Identifying General Patterns of Emerging Topics

A total of 55,980 models for 311 domain topics over 20 years were trained with nine sets of training data using the LR algorithm. The general patterns are first extracted using the whole 55,980 models, averaging the coefficients and intercepts from multiple models to calculate the averaged linear predictor $lp = x^T\beta + \beta_0$. Using the average coefficients and intercepts for *all* models, the linear predictor equation using all 15 features with $4dp$ becomes as follows:

$$\begin{aligned} lp_{15} = x^T\beta + \beta_0 = (&-0.1407\#nodes - 472.0608cohesion \\ &+ 5.9879density + 1.2630transitivity - 0.0034norm.tri \\ &- 1.0100avg.path + 41.4201avg.pr + 6.0566avg.dc \\ &+ 10.4910avg.btc - 0.0229avg.age - 0.0019\#edges \\ &- 0.3995avg.deg - 0.0008avg.deg.w + 0.0024avg.edge.w \\ &+ 28.4043avg.c.w) - 1.8725 \end{aligned} \tag{2}$$

The logistic equation is applied to generate the binary outcome $predicted = \lfloor e^{lp_{15}}/(1+e^{lp_{15}})\rceil$. The *predicted* labels are then compared to the actual data.

Figure 3 shows the prediction results of the generalized linear predictor equation lp_{15} over 311 domain topics over the x-axis, ordered by their F1. A majority of the domains maintained relatively high F1, indicating that there is a common pattern across the majority of the research domains when it comes to the prediction of new topic emergence. F1 below 0.5 in eight domains shows that there are indeed domains where a common pattern does not apply or even results in reversed predictions. This is expected as there are various research domains with different research conventions; there will be domains with unique research behaviors, unlike the others. This is supported by the fact that the five d with the lowest F1 are all medical-related domain

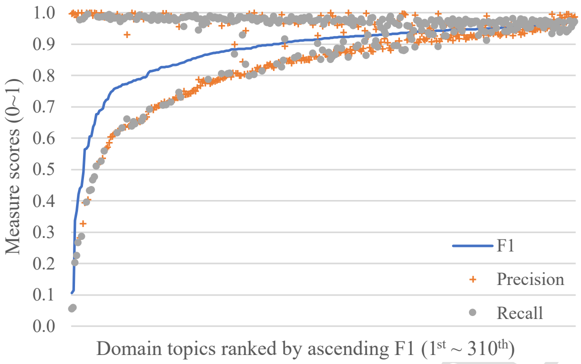

Figure 3: Precision, Recall, and F1 of the topic emergence prediction using the generalized linear predictor equation lp_{15} over all $T_{d,y,500,9}$, sorted by F1 in ascending order.

topics (*endocrinology, internal medicine, surgery, medicine, and cardiology*), and three out of five in ranks 6 to 10 are related to the specific fields of the economic domain (*Keynesian economics, classical economics, and neo-classical economics*). This is not necessarily the result of the small dataset size; individual classifications done in the previous section showed on average 0.9987 and 0.8436 *auc* for medical and economics-related topics. The topics were individually able to predict topic emergence, albeit less so with economic-related topics due to them being the three fewest topic groups.

The topics with lower prediction accuracies have distinct differences in their low F1; medical-related topics show low recalls, while economics-related topics show low precisions. Medical-related topics in Table 4 show larger average values for features with negative coefficients in lp_{15}; *avg.deg.w* is 8.6109 times larger than that of the top 10 domain topics, for example, reflecting a very high topic co-usage pattern in the domain compared to the rest of the research fields. Differences in the feature values result in the linear predictor outcome being reduced by 16.7186 on average, which leads to more *False* predictions and lower recall. Economics-related topics on the other hand show an average increase of 9.4488 in the linear predictor outcomes leading to more *True* predictions with lower precision. This is mainly contributed by higher *avg.pr* as well as lower *avg.deg* and *avg.deg.w*, indicating that these domains share fewer topic co-occurrences.

Several partial lp were generated from different sets of models to identify a common pattern specific to the given subset. $lp_{l,n}$ are first generated to

Table 4: Differences between Five Features for Low-performing Five *Medical*, Three *Economic* Topics, and Ten Topics with the Highest F1 using lp_{15}.

Features	Medical	Economics	Top 10
norm.tri	131.9102	6.2050	57.8298
avg.pr	0.0574	0.1688	0.0812
#edges	314.1599	17.3343	140.6523
avg.deg	14.9470	3.9370	9.6520
avg.deg.w	24654.3094	346.7999	2863.1532

represent common patterns specific to the l and n combination; they resulted in statistically insignificant F1 values with $p = 0.1494$ when predicting the topic groups within the combination; there are no significant differences in predictive capability with different training data sizes. Extracting the linear predictor for a specific FoS level resulted in significant differences on the other hand. $lp_{level=0}$ and $lp_{level=1}$ were generated by averaging the coefficients and intercepts of the models derived from FoS with the given level, each representing the common pattern in the top two levels of the six-level topic hierarchy. Figure 4 shows the prediction results with *from.level_to.level* as columns; when they are applied to the respective FoS they were averaged from (0_0, 1_1), as well as the prediction results for different levels (0_1, 1_0). The lp_{15} result for all FoS is shown as All_All. The first noticeable changes are in 0_0, having the highest mean without a long tail of outliers. This indicates that the high-level domains share a common pattern detectable with high resiliency. This overarching common pattern within the level 0 FoS resulted in fewer severe prediction failures in 0_1 and 1_0 as well, either when it was captured during the training or when it was identified during the testing. The lower quartile values for 0_1 indicate the common patterns found in the lower level are harder to predict using the higher-level topics; while the level 0 topics share a distinct common research pattern, more focused level 1 topics have more variance in their common patterns.

Analysis on the effect of feature filtering revealed that the main contributor to the poor performance in the majority of worst-performing models is the *avg.deg.w*, the mean frequency-weighted degree. 13 out of the bottom 20 topics showed a sharp increase in F1 when it was removed from the linear predictor equation, reaching F1 > 0.9. Six out of seven that did not experience a significant increase in F1 (*Keynesian economics, classical economics,*

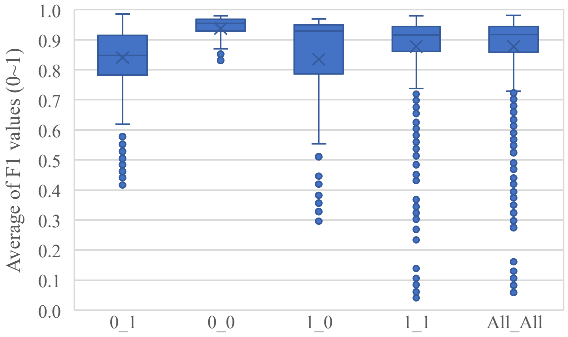

Figure 4: F1 with FoS level hierarchy information. $lp_{level=0}$ and $lp_{level=1}$ are applied to the domain topics of the same level (0_0, 1_1) and different level (0_1, 1_0).

Table 5: F1 with Iterative Feature Removal, with the Number of Remaining Features #.

Removed	#	F1	Removed	#	F1	Removed	#	F1
None	15	0.8770	*avg.path*	10	0.9236	*avg.btc*	5	0.8852
avg.deg.w	14	0.8951	*avg.c.w*	9	0.9242	*avg.age*	4	0.9141
density	13	0.9237	*transitivity*	8	0.9150	*cohesion*	3	0.8358
norm.tri	12	0.9241	*#nodes*	7	0.9221	*avg.dc*	2	0.7783
#edges	11	0.9244	*avg.edge.w*	6	0.9196	*avg.deg*	1	0.7729

neoclassical economics, polymer science, astrobiology, and *earth science*) can be explained by their small dataset sizes, leaving only *biochemical engineering* unexplained. A sharp increase with lp_{14} shows that the frequency of topic co-occurrence resulted in an adverse effect when building a general predictor model. More features were removed from lp_{15} to identify the adverse effect each feature has on the prediction outcomes while generating a shorter linear predictor equation with performance similar to the full equation in (2).

Table 5 shows the prediction performance with a diminishing number of variables in lp_{15}, iteratively applying the removal of a variable that results in the highest F1. The order in which the features were removed has a moderate correlation with feature importance with a correlation coefficient of 0.4679, showing that there were features with higher importance that were more detrimental to the performance. The F1 increases with the removal of the first three features *avg.deg.w*, *density*, and *norm.tri*, reaching up to F1 =

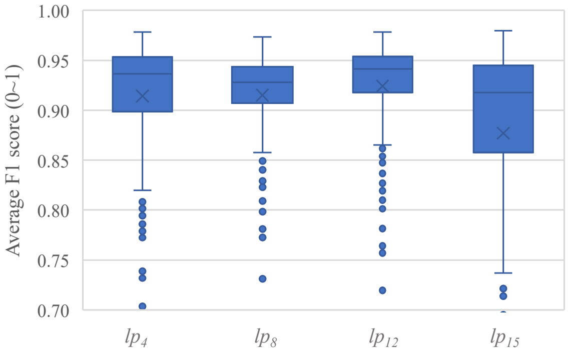

Figure 5: Comparison between lp_{15}, lp_{12}, lp_8, and lp_4 using box and whiskers plots of their F1 distributions. The features are removed following the order in Table 5, and the y-axis is cut out at F1=0.70 removing a few outliers from the figure.

0.9244. The predictor then retains most of the prediction performance with a decreasing number of features, keeping the F1 value above 0.9 until only four of the features were used except for one outlier with five features. The resulting general logistic regression model and the shortened linear predictor equation lp_4 with the four remaining features are

$$\begin{aligned} predicted &= \lfloor e^{lp_4}/(1+e^{lp_4})\rceil, \text{and} \\ lp_4 &= (-472.0608 cohesion + 41.4201 avg.pr \\ &\quad + 6.0566 avg.dc - 0.3995 avg.deg) - 1.8725 \end{aligned} \tag{3}$$

which can predict topic emergence related to 311 individual domain topics with an average F1 > 0.9 with the lower quartile value of 0.8988 (4dp).

The equation calls for a lower cohesion and average degree, and higher PageRank and degree centrality for the *new* labels; weaker within-connections are desired. The shortened equation contains features from subgraph, node, and edge-related categories but not the weighted category; while the information external to the topic group is needed, the node and edge frequencies are not required. Figure 5 shows that while the predictors share a similar maximum F1, predictors with fewer features have higher performance when the worst three features were removed. Reducing the number of features from 12 to 4 increased the performance variance but retained F1 > 0.9 for nearly 3/4 of the datasets. This shows that the shortened linear predictor does contain knowledge about the emergence of new topics in terms of their future ancestors and their subgraph properties in the topic network.

6. Conclusion

The network-based topic emergence prediction method is applied throughout the wide range of research domains in this paper to validate the generalizability of the method and to identify a general pattern for emerging research topics. The emergence of a *new* topic is defined as a node added to a network. It is assumed that the shared interest matures over time until a new topic is formed as a common neighbor, and the materialization of such topics is reflected in specific structural features of their neighboring topics within the topic network. This approach allows the prospective prediction based on the observed patterns by simple means of extrapolation, which is harder to achieve with the more traditional text-based topic models without having access to the non-existent future documents. Topic co-occurrence patterns in different domains were first captured by applying multiple binary classification models to show the method is generalizable with different ML algorithms over a variety of research domains. The trained models are then averaged to create a general predictor equation classifying whether the future neighbor of the given topic subgraph is *new*.

The binary classifiers were able to distinguish ancestors of *new* topics with high accuracy. Barring a few outliers with insufficient data, the performance retained most of its values even with the lower number of rows used. Training the model with only one topic network in the previous year resulted in *auc* and *aucpr* over 0.98, showing a clear tendency of having linear data by having the logistic regression as the most frequently best-performing model. The high performance is pervasive over the majority of the datasets, each centered around one of the domain topics, showing *auc* over 0.9 for 306 out of 311 datasets. The results show that the new topics exhibit distinct, measurable differences from the existing ones, and the existing ML algorithms can be trained to identify the emergence of new topics without optimizing individual tasks with hyperparameter tuning.

The logistic regression algorithm has the highest model interpretability, and the trained models are analyzed to evaluate the possible existence of a common pattern for emerging topics. The results showed that there is indeed a domain-independent pattern that can predict topic emergence with high accuracy. The averaged predictor resulted in F1 above 0.9 for 59.68% of the experimented domains. The low performances in some of the medical and economic-related domains indicate that they have distinctive topic group structures, which can be shared among a specific subset of research fields.

A universally applicable predictor model will be available with hierarchy or similarity-based model grouping.

The results showed that the topic group data not only are linear but also have lower dimensions than previously assumed. Removing features resulted in heightened performances, resulting in the generalized topic predictor function with four features. The proposed function accurately predicted topic emergence in a variety of academic domains, validating the assumption that emerging topics share a general pattern. Future work can include extracting a set of topic group candidates without accessing the future neighborhood information, which would allow prospective topic evolution prediction.

References

[1] S. Jung, A. Segev, Analyzing the generalizability of the network-based topic emergence identification method, Semantic Web 13 (3) (2022) 423–439. doi:10.3233/SW-212951.

[2] A. Gohr, A. Hinneburg, R. Schult, M. Spiliopoulou, Topic evolution in a stream of documents, in: Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM, 2009, pp. 859–870.

[3] N. Ilhan, Ş. G. Öğüdücü, Feature identification for predicting community evolution in dynamic social networks, Engineering Applications of Artificial Intelligence 55 (2016) 202–218.

[4] Y. Shao, X. Li, Y. Chen, L. Yu, B. Cui, Sys-TM: A fast and general topic modeling system, IEEE Transactions on Knowledge and Data Engineering (2019) 1–1doi:10.1109/TKDE.2019.2956518.

[5] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 113–120.

[6] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005, pp. 198–207.

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

[8] B. Chen, S. Tsutsui, Y. Ding, F. Ma, Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval, Journal of Informetrics 11 (4) (2017) 1175–1189.

[9] M. Menenberg, S. Pathak, H. P. Udyapuram, S. Gavirneni, S. Roychowdhury, Topic modeling for management sciences: A network-based approach, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 3509–3518.

[10] Y. Jo, J. E. Hopcroft, C. Lagoze, The web of topics: discovering the topology of topic evolution in a corpus, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 257–266.

[11] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, L. Giles, Detecting topic evolution in scientific literature: how can citations help?, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 957–966.

[12] L. Dietz, S. Bickel, T. Scheffer, Unsupervised prediction of citation influences, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 233–240.

[13] H. C. Ozmutlu, F. Çavdur, Application of automatic topic identification on excite web search engine data logs, Information Processing & Management 41 (5) (2005) 1243–1262.

[14] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A survey on knowledge graph-based recommender systems, IEEE Transactions on Knowledge and Data Engineering (2020) 1–1doi:10.1109/TKDE.2020.3028705.

[15] A. L. Porter, M. J. Detampel, Technology opportunities analysis, Technological Forecasting and Social Change 49 (3) (1995) 237–255.

[16] C. Balili, A. Segev, U. Lee, Tracking and predicting the evolution of research topics in scientific literature, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 1694–1697.

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

[18] S. Jung, R. Datta, A. Segev, Identification and prediction of emerging topics through their relationships to existing topics, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 5078–5087.

[19] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701 (2012).

[20] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 625–632.

[21] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[22] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, R. Rogahn, A review of microsoft academic services for science of science studies, Frontiers in Big Data 2 (2019) 45.

[23] S. E. Hug, M. Ochsner, M. P. Brändle, Citation analysis with microsoft academic, Scientometrics 111 (1) (2017) 371–378.

[24] Z. Shen, H. Ma, K. Wang, A web-scale system for scientific knowledge exploration, arXiv preprint arXiv:1805.12216 (2018).

[25] Y. Su, Y. Zhang, D. Ji, Y. Wang, H. Wu, Ensemble learning for sentiment classification, in: Workshop on Chinese Lexical Semantics, Springer, 2012, pp. 84–93.

Credit Author Statement

CRediT Author Statement

Sukhwan Jung: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing - original draft

Aviv Segev: Conceptualization, Formal analysis, Investigation, Methodology, Writing - original draft , Writing - review & editing

Highlights

Identifying a Common Pattern within Ancestors of Emerging Topics for Pan-Domain Topic Emergence Prediction

Sukhwan Jung, Aviv Segev

- Content transition within a topic and correlation between topics are distinguished.
- Generalizability of the method is validated over a wide range of research domains.
- Pan-domain features are identified covering various topic co-occurrence behaviors.
- High percentages of emerging topics were detected using logistic functions.
- The general linear predictor equation with four features showed an average F1 > 0.9.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: