Analyzing the generalizability of the network-based topic emergence identification method

Sukhwan Jung[a,*] and Aviv Segev[a]

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

Abstract. The field of topic evolution helps the understanding of the current research topics and their histories by automatically modeling and detecting the set of shared research fields in the academic papers as topics. This paper provides a generalized analysis of the topic evolution method for predicting the emergence of new topics, where the topics are defined as the relationships of its neighborhoods in the past, allowing the result to be extrapolated to the future topics. Twenty fields-of-study keywords were selected from the Microsoft Academic Graph dataset, each representing a specific topic within a hierarchical research field. The binary classification for newly introduced topics from the years 2000 to 2019 consistently resulted in accuracy and F1 over 0.91 for all twenty datasets, which is retained with one-third of the 15 features used in the experiment. Incremental learning resulted in a slight performance improvement, indicating there is an underlying pattern to the neighbors of new topics. The result showed the network-based new topic prediction can be applied to various research domains with different research patterns.

Keywords: Topic Evolution, Topic Prediction, Network-based Topic Modeling, Scientometrics

1. Introduction

Scientific knowledge evolves through the contribution of researchers around the globe; discoveries are made to expand the existing research topics or to contribute towards the creation of new topics. The gradual expansion or transition of research topics based on the foundation of past knowledge guarantees the validity and soundness of the research. Identifying and predicting the emergence of new topics are therefore dependent on understanding the related topics representing the set of shared themes, or research fields. They can appear in various forms, including the philosophical category of the research, theoretical development of research models, applications of the technology, and specific algorithms. Identifying such topics in the academic papers is therefore a crucial part of research activity. Researchers understand the topics by first reviewing a multitude of articles, internalizing the evolution occurring within the researchers' fields of interest, which in turn allows them to ascertain the desirable paths the current and future research can take. A better understanding of such knowledge allows more targeted research aimed at highly demanded topics, which is needed in both academic and industrial fields.

Traditional topic evolution methods mimic the process by utilizing text-based topic models to understand the topic in each document collection and track topical changes over time. Topic modeling methods extract statistical constructs based on word co-occurrences in the given document collection, where changes in topics can only be measured by the differences between the content of two topics; connections and correlations between different topics are not incorporated into the traditional topic modeling methods [21]. Topic evolution methods are therefore mostly limited to identifying content transition within a given topic, not how it is correlated to other topics. Unforeseen topics in the future cannot be modeled without having access to the set of future documents

[*] Corresponding author. E-mail: shjung@southalabama.edu.

yet to be written. As a result, topic evolution based on traditional topic modeling methods is not suited to predict new topics.

A previously proposed network-based approach identified emergence on topic networks where the definition of topics based on its neighbors' previous relationships intuitively allows extrapolations to future new topic predictions [18]. The definition allowed the topics in a certain timeslot to be classified based only on the structural data available in previous timeslots, showcasing a novel functionality of predicting topic evolutions solely with the topic co-occurrences using journal-specific publications as the dataset. This paper expands on this research by testing the generalizability of the method, offering a better understanding of the network-based topic prediction method.

The goal of the proposed method is to capture the emergence of new topics, which can be explained by their correlation to the existing topics. This can be formalized as classifying subgraphs in the given topic network as to-be-neighbors of new topics in the future based on their graphical properties. The topic networks are first extracted from an open bibliographical dataset, with each network representing publications in a specific research journal with a focused set of research interests. The topic network is divided yearly to generate an evolving network, where each topic in timeslot *y* is either *new*, appearing for the first time in *y* for the given topic network, or *old*. A binary machine learning algorithm is trained using the neighbors of each node in the previous years, classifying the neighbor subgraphs in the past having *new* or *old* topics as their future neighbors. Twenty topic networks were generated from publications related to twenty highly used fields-of-studies from the Microsoft Academic Graph[1] dataset. The impact of different features and the number of features impactful to the classification performances are analyzed. The topic co-occurrence patterns representing new topics in the scientific bibliographic records are incrementally learned over time within a single dataset to capture domain-specific knowledge and their evolutions. The same process is then tried over different datasets to capture underlying common patterns throughout the different knowledge domains. The experiment results showed that the proposed method retains its high classification accuracy with all 20 datasets with less than one-third of the 15 features while showing relatively small, but statistically significant, performance improvement using the incremental learning.

Section 2 reviews the related work on topic evolution, previous attempts on the prediction of new topics, as well as background research for the proposed method. Section 3 and 4 detail the proposed method and experimentation, and the experiment results are shown in Section 5.

2. Related work

2.1. Identifying the evolution of topics

Automatically identifying topical changes within the document set requires methods to extract machine-readable topics from the collection. Topic modeling provides a statistical approach to discover *topics* within a given corpus, where topics are modeled as the latent semantic structures in the form of word-popularity sets based on the statistical distribution and word co-occurrences.

Latent Dirichlet Allocation (LDA) [5] finds latent topics within a document collection and is one of the most widely used topic modeling methods on which many other methods are based [13,22]. Word-topic links are iteratively assigned with word co-occurrences between documents; topics, defined as word distributions over a corpus dictionary, are then assigned to each document [33]. Topic evolution aims to identify the evolution of such topics in a sequentially ordered document collection. Document collection is first divided either uniformly or irregularly [12] into sequentially-ordered sub-collections on which topic models independent of the neighboring sub-collections are generated. Temporal topic models are then connected over time with similarity measures, and changes in the topics are sequentially analyzed to identify the evolution of topics.

Dynamic topic models [4] are one of the early implementations of topic evolution, focusing on capturing the changes within a set of chained topics with fixed timeslots where the Kalman filter and wavelet regression are used to approximate natural parameters of the topics found at different time slices. Evolutionary theme pattern mining has tried to capture not only the changes within each topic but also the sequential connections over multiple topics [24]. The Kullback-Leibler divergence is used as a distance metric between topics, and the topics on different timeslots are designated as having an evolutionary transition when their distance stays below dataset-

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

specific thresholds. The collection of such evolutionary transitions results in detecting merge and split events over time as multiple connections are allowed between different topics. A similar approach is made by utilizing cross-citations between topic pairs' member documents as well [16].

Topic evolution in conjunction with bibliographical dataset analysis has been tried by numerous researchers to better identify the topic evolution events. The citation contexts are used in an iterative topic evolution learning framework to increase the performance of topic evolution with better topic models [14], where the document collection is expanded by the documents cited by its members. The inheritance topic model [9] is utilized to classify papers into autonomous parts with originalities and parts inherited from cited documents. Differentiating two parts allowed the method to overcome the topic dilution with cited papers, generating more new topics compared to LDA-based approaches.

A more recent approach to topic evolution utilizes communities of keywords in a dynamic co-occurrence network [2]. The medical subject headings dataset from PubMed[2] was used to build a filtered co-occurrence network of major subjects within the medicine domain divided into five-year snapshots. Word clusters were found and linked to generate the evolution of topics over time. Topic evolution based on two-tier topic models is tried for a better merge and split detection, where topic correlations in the same timeslot are used to identify topic evolution [7]. Timeslot-specific local topics are extracted from yearly divided sub-collections of documents, while time-spanning global topics are retrieved using the whole corpus. Global topics stay static, having connected to dynamic local topics at each timeslot with cosine similarities above a given threshold. Changes in the number of local topics connected to global topics are then used to define the topic evolution events; decreased and increased numbers of local topics connected to a global topic respectively represent merging and splitting of the topic.

2.2. Identifying and predicting new topics

Topic Detection and Tracking (TDT) [10] aims to capture the appearances of new topics in continuously generated text data in real-time; a topic is defined as "*a seminal event or activity along with all directly related events and activities*" [10]. First story detection (FSD) is one of the parts of TDT research tasks. The goal of FSD is to search and organize new topics from multilingual news articles, or identifying the first article introducing the new story [1]. Topic-conditioned FSD with a supervised learning algorithm first classified news articles into a set of predefined topic categories before identifying novelty within each topic [39]. FSD is also used in conjunction with document clustering to identify the earliest report of a certain event in news articles [35].

Identification of emerging topic trends has led to the division of research front and intellectual base, where the latter is an established foundation of domain knowledge on which the former is built. The underlying assumption is that the citation and co-citation between articles transfer the existing knowledge from the intellectual base to the research front. The CiteSpace II [8] further utilized a keyword co-occurrence relationship by employing a bipartite graph of keywords and articles. Research front terms are identified by the sharp frequency growth, and then used to identify research front articles, which in turn are absorbed into the intellectual base in the next time slice. Burst term detection, in conjunction with keyword co-word analysis, allows multi-dimensional exploration of the research front in question [23].

While these approaches allow the detection of merging and splitting of time-spanning topics and their transitional ratio at the temporal level, the use of the text-based topic models inherently limits the predictive capabilities; the evolutionary events such as emergence, merge, or split can only be retrospectively analyzed once the topic is captured from the document set. Using author groups from a bibliographic dataset for determining topics connected over time by authors showed that when topics defined by the authors are used instead of NLP-based topic models, topic evolution on the temporal network is possible; the topic evolution events are defined by the network structures and therefore a predictive analysis is possible [21].

On top of the emergence events detected by the appearance of topic models dissimilar to the ones in the previous timeslots, there are a number of research studies dedicated to identifying new topics with a varying definition of the topic. One such field is new topic identification, where the topic is defined as the entities the user is interested in during the search engine querying session; the query patterns and the intervals between queries are used to identify topics [26]. Neural network (NN) is introduced to reduce the errors in new topic estimations based on typos by utilizing the character n-gram method to bypass spelling errors in the queries [11]. There are also sev-

[2] https://www.ncbi.nlm.nih.gov/pubmed/

eral researches focusing on utilizing the queries' statistical characteristics, such as search patterns, frequency of queries, and the relative position in the querying sessions [27].

Technology forecasting [28] is another field of research aiming to predict the characteristics of technology in the future; the technology, or topic, is defined as a representative keyword instead of a statistical model. Various techniques from simple extrapolation to organization management [3] and fuzzy NLP [25] are used to identify and predict changes in technology indicators [6]. Multiple applications of the predictive topic evolution have been proposed, including a semi-manual technology trend analysis which was done to identify the roots of new technologies with their projected impact on the research field [29].

A previously proposed technology trend analysis approach with multiple data sources shows that while different data sources exhibit different forecast speeds, predicting the growth and shrinking in technology trends is possible extrapolating on a previously known technology growth curve [30]. A network-based approach was proposed to overcome the rigidity of trend-based forecasting where the prediction is dependent on the type and shape of the technology growth curve used. Node prediction based on preferential attachment link prediction is proposed to classify whether the nodes in citation networks have a connection to a new node in the future [19], labeling the new nodes by utilizing the metadata of their neighboring nodes [20]. This showed that predicting nodes in bibliographic networks is possible based on the structural properties of the network. More complex contexts of the new nodes in knowledge networks were extracted by identifying the neighbors of the new node in the past timeslot to formulate the context of the new node solely based on the metadata of its to-be-neighbors [17].

Network-based topic emergence identification is a network-based approach to a new topic prediction by utilizing a topic network. The emergence of new topics was identified by capturing the relationships between their neighborhoods in the previous years, and predictions based on the existing clustering algorithms were made to validate the possibility of proactive topic emergence predictions with the proposed method. This paper aims to show the generalizability of the proposed method using various datasets with different focus and interests, capturing the shared knowledge between knowledge domains with an incremental learning method to improve the performance.

3. Network-based new topic identification and prediction

3.1. Generating topic networks

NLP-based topic modeling can be used on the document collection dataset to retrospectively identify topics already present in the research field but has limited capability to prospectively predict the appearance of previously unused topics in the future without the documents to extract topics from. The proposed method utilizes a topic network instead, where emerging topics in a bibliographic dataset equate to new nodes in the topic network. Textual metadata is not considered for analysis, and only graphical structures are used.

The topic network $T_y = (V, R_y)$ represents co-occurrence frequencies R_y between topics V within the knowledge domain at given year y. Topic set V consists of the topic node u and the year the topic is first used in the dataset fy, and R_y is the weighted edge set between nodes u and v, with w_y as co-occurrence frequencies in y.

$$T_y = (V, R_y), \textit{ and } V = (u, fy)\ R_y = (u, v, w_y) \quad (1)$$

3.2. Classifying subgraphs by the common neighbors

The proposed method is run on the topic network T_y in Eq. (1), where topics in year y are classified as *new* or *old* based on the structural features of their neighbors. Neighborhoods *neighbors(v, y)* of each topic v in year y are extracted to build a set of neighborhoods N_y from T_y. Each neighborhood is then categorized into two groups by the age of v calculated by $fy(v) – y$ categorizing whether the topic v first appeared in the given year y, in which case $fy(v) = y$. The state of v, $C(v)$ is calculated as the ceiling of topic age normalized by the oldest topic, where the *new* topics are denoted by $C(v) = 0$. Any preexisting topics have non-zero ages, and the normalized ceiling function result in $C(v) = 1$.

$$N_y = \{neighbors(v, y) \mid v \in V_y\}, \text{ and}$$

$$C(v) = \lceil (fy(v) – y) / (\max(fy(V)) – y) \rceil \quad (2)$$

More prominent topics are likely to co-occur with more topics, and therefore the top 100 topics with the largest number of nodes in N_y are selected for each label $C(v) = 0$ and 1, resulting in a total topic count of 200 for each classification task. In case the number of instances for one label is below 100, then the number of v for the other label is reduced further to have the same number of instances for both labels.

Evolution of existing topics such as merge and split is not targeted, and hence there is no need to train the classifier for the gradual evolution events within existing topics. Temporal features are therefore not analyzed; only static features are used in the experiment. Table 1 shows the list of 15 structural features of the neighbor subgraphs used to train the binary classifiers. These features characterize the subgraph quality in several aspects and are grouped by the component they are used to measure, including six properties related to the whole subgraphs, four average values of member node properties, two properties related to the number of edges, and three properties weighted by the topic co-occurrence frequencies.

3.3. Classifying new topics with incremental learning

The emergence of new topics is the only event being searched; therefore the binary classification on year y is trained by neighbor subgraphs in previous years. Sets of open neighborhoods $Train_{y,t}$ and $Test_y$ are generated using t previous topic networks. The same set of neighbors $n = neighbors(v)$ is used to identify open neighborhood subgraphs of v in multiple previous timeslots, denoted by $T_k(n)$ where $y\text{-}t \leq k \leq y$.

Table 1 Structural features used in the experiment.

Features used	Description
Subgraph	
Node Count	Number of nodes
Cohesion	Number of internal/external edges
Density	Number of observed/possible edges
Transitivity	Number of observed/possible triangles
Normalized Triangles	Number of triangles/nodes
Mean Shortest Path	Mean of all node pairs’ shortest paths
Nodes	
Mean PageRank	Mean PageRank for subgraph nodes
Mean Degree Centrality	Mean degree centrality for subgraph nodes
Mean Betweenness Centrality	Mean betweenness centrality for subgraph nodes
Mean Node Age	Mean age for subgraph nodes
Edges	
Edge Count	Number of edges in the subgraph
Mean Degree	Mean degree in the subgraph
Weighted	
Mean Degree Weighted	Mean degree with edge weights
Mean Edge Weighted	Mean edge weights
Mean Clustering Coefficient	Mean weighted clustering coefficient

$$sub(v, y, k) = \{(n, \{n_i, n_j\}) \mid n \in neighbors(v, y), \{n_i, n_j\} \in E_k \},$$

$$Train_{y,t} = \{\mathrm{sub}(v, y, y\text{-}t) \cup \ldots \cup sub(v, y, y\text{-}1) \mid v \in V_y\}, \text{ and}$$

$$Test_y = \{sub(v, y, y) \mid n \in N_y\} \tag{3}$$

Neighbor subgraphs in Eq. (3) represent interactions within direct predecessors of *new* topics and neighbors of preexisting *old* topics, which are shown to have distinguishable structural features in the previous research [18]. The classification accuracies, precision, recall, F1, and area under the ROC curve (AUC) based on subsets of 15 features are compared to show the effect of the number of features as well as the features with the most importance.

The proposed method trains a machine learning algorithm to classify new topics by past interactions within their neighborhoods for a given knowledge domain. The generalizability of the proposed method is analyzed by implementing an incremental learning approach. The trained model is retained for each of the incremental learning instead of being re-initialized. *Within-domain* learning is done over incrementing y within each of the knowledge domain to incrementally adapt to the continuous topical interactions over time. *Between-domain* learning is done between each of the domain pairs at the same y resulting in a total of $k \times (k\text{-}1)$ pairs for k number of domains used in the experiment, aiming to test the possibilities of incremental learning between different knowledge domains. Changes in its performance, when two domains share the same parent domain, are observed as well. Increases in the performance would suggest that the topic networks at different times and under different domains share underlying models. The proposed method would then be generalizable to any parts of the knowledge stored in the bibliographic records.

4. Experiments

4.1. Dataset preprocessing

Multiple topic networks were generated from bibliographic records extracted from the Microsoft Academic Graph (MAG) [34], which is a heterogeneous bibliographic dataset [32]. The MAG is selected as the source dataset for two reasons. Firstly, it was deemed competitive with major bibliographic search engines such as Google Scholar or Scopus, even with relatively recent creation [15]. Secondly, the MAG has a built-in ontology called fields-of-study (FoS) representing each paper with different hierarchical concepts [31]. A six-level hierarchy of concept is generated each month using knowledge base type prediction with Wikipedia articles, employing graph link analysis and convolutional neural network methods. The hierarchical concepts are then tagged to the papers using a large-scale multi-level text classification method on pre-trained word embedding vectors. The tagging is done weekly to keep up-to-date concept assignments. Identifying dataset-wide topics in a large-scale dataset is by itself a huge task; therefore the tagged FoS are defined as the topics for the document in this paper. While the author-assigned keywords in research publications also represent their topics, the MAG database does not have keywords as one of its relational database tables and therefore is not used in the experiment.

The MAG dataset snapshot in February 2020 is downloaded for preprocessing through the Microsoft Azure Databricks, containing 197,642,464 publications, 709,934 FoS, 48,829 journals, more than 1.5 billion citation links, and 1.3 billion paper-FoS links. Analyzing the whole graph would be too complex to compute, and therefore data subsets are extracted as the bibliographic records related to selected FoS, each representing subsets of topics focused on different research fields.

Table 2 Twenty FoS in the February 2020 MAG dataset.

Rank	DisplayName	MainType	Lv
9863	usability	business.industry	2
9299	software development	business.industry	3
8335	polysaccharide	chemistry.chemical_classification	2
8494	hydrogen peroxide	chemistry.chemical_compound	2
8442	ozone	chemistry.chemical_compound	2
8868	palladium	chemistry.chemical_element	3
8480	cadmium	chemistry.chemical_element	3
9749	diamond	engineering.material	2
9216	drainage basin	geography.geographical_feature_category	2
9961	calcination	law.invention	3
8177	fertility	media_common.quotation_subject	3
9058	unemployment	media_common.quotation_subject	2
9964	physical examination	medicine.diagnostic_test	3
8153	malaria	medicine.disease	3
8349	thrombosis	medicine.disease	3
7579	air pollution	medicine.disease_cause	2
9171	activated carbon	medicine.drug	3
12641	saline	medicine.medical_treatment	3
9418	stent	medicine.medical_treatment	3
12338	gaussian	symbols.namesake	2

The FoS are selected by the following criteria. The size and activity of the datasets are modulated by selecting FoS with 100,000 < related publication count < 120,000 and 1,000,000 < combined citation count < 1,500,000. FoS without the main type data are removed to ensure that each dataset's parent domain is known, selecting two FoS from each main type with the highest ranking.

Table 2 shows the resulting 20 FoS with ranks measured by the possible importance along with the display name of the FoS, their main type within the FoS hierarchy, and the level of the FoS in the hierarchy tree. 20 FoS-specific datasets are extracted into the SQL databases using a high-performance computing service by Alabama Supercomputer Authority[3]. All data rows in the *PaperFieldsOfStury* table containing the matching *FieldOfStudyId* are retrieved, then rows matching the filtered papers in *PaperFieldsOfStudy* and *FieldsOfStudy* tables are retrieved for FoS used in the journal and how they are assigned to individual publications. With a series of SQL queries, *FirstUsedYear* column is added to *FieldsOfStudy* tables to represent the first year *fy* the given FoS is used with the FoS, and *FOSneighborCount {Node1, Node2, Year, Frequency}* table is created to represent undirected links within each dataset with node pair *u*, *v*, year *y*, and frequency *w*, where FoS are the nodes and the links represent the two FoS assigned co-occurring in the same publications. *Frequency* shows the co-occurrences between two FoS, which is divided for each *year* to distinguish different FoS links and weights at different years.

4.2. Generating topic networks

After the dataset preprocessing is done, the topic network T_y in Eq. (1) for each FoS is generated for y=[2000,…,2020]. The year y is ranged to retrieve the detection of newly used topics in the 21st century. For each FoS, SQL queries are run on the *FOSneighborCount* table to extract topic co-occurrence with *FOSneighborCount.Year = y* where the *Year* column in the *FOSneighborCount* table represents the year the topics co-occurred. The resulting edge data R_y is used to build a topic network using the equation in Eq. (1).

4.3. Classifying subgraphs by the common neighbors

Data downsampling is done on each dataset with *C(v)* as the class variable. This is done to reduce the

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

total amount of data while balancing the number of labels for the classification. Isolated nodes are ignored as there are no neighbors to analyze. Data standardization is also done to remove range differences between 15 features, where the values of each feature are first subtracted by the average value and then divided by its standard deviation.

$$z = (x - \mu) / \sigma \quad (4)$$

Training size is set to *t=9* as the increase in the classification performance diminishes with large *t* values. Initial experiments showed the Logistic Regression (LR) was one of the best performing algorithms without showing anomalous classification patterns over combinations of classification variables. The L-BFGS algorithm [36] is used as an optimization function for the ML model, with a maximum training iteration of 100.

4.4. Classifying new topics with incremental learning

Feature selection is done for all feature count *f* = 1,…,15. For each *f*, combinations of features with length *f* are compared by different score functions shown in Table 3, utilizing f-values and mutual information of the classification results. To analyze the importance of the features, one classification model is trained using the *selected* features while another is trained using the *excluded* features. 2-dimensional principal component analysis (PCA) is also done to test the linear separability of the features.

Table 3 Descriptions of four score functions used in classification.

Score Function	Description
f_classif	ANOVA F-value between labels
f_regression	F-value for univariate regression
mutual_info_classif	Estimated mutual information between labels
mutual_info_regression	Estimated mutual information for continuous target

Incremental learning is implemented in two different ways, named after the function names they are based on. The *warm* approach retains the coefficients of the trained model which are used as the initial coefficient in subsequent training, while the *partial* approach incrementally trains the model with additional data. *Sklearn* Python library's *warm_start* attribute and *partial_fit* function are used respectively. Both approaches have limitations; the *warm* approach risks overwriting the initial training result when there are major shifts in the new training data, while the *partial* approach would suffer performance losses in such case as it would try to search for the solution covering both datasets. These are compared against the non-incremental *cold* approach, where the training model is re-initialized before every training.

To analyze the possible differences between different classification algorithms, a linear support vector machine (SVM) algorithm is used in addition to the logistic regression used in the previous section. Different *epoch* values are tested to show the effect of the epoch sizes. The *partial_fit* function only trains the model one generation at a time while the model with *warm_start* attribute is trained over multiple epochs; hence it is repeated *epoch* number of times to mimic the incremental learning with multiple epochs. The number of data rows affects the incremental learning performances; hence a different number of topics is also tested.

$$\text{epochs} = [10, 50, 100], \text{ and}$$

$$\text{num\_topics} = [10, 50, 100, 200]^{a} \quad (5)$$

[a]200 for between-domain.

Within-domain learning is done over *y* from 2000 to 2019 for each FoS dataset testing the incremental adaptation to the continuous topical interactions over time. *Between-domain* learning is done between each of the FoS pairs at the same *y*, instead. This results in a total of 380 pairs for FoS used in the experiment; only topic=200 is used for *between-domain* learning with *y=[2000,2005,2010,2015]*.

5. Results

5.1. Classifying new topics using FoS datasets

The classification results were measured excluding *y=2020* as the performance is significantly lower for all FoS datasets in the last year with Acc = 0.4068, AUC = 0.8028, and F1 = 0.5589. This is because the MAG dataset used in the experiment has only partial records of the 2020 publications up to February. This is supported by the retained high recall value for *y=2020*. The model failed to distinguish between new and old topics based on their incomplete neighborhoods, classifying all candidates as a single label resulting in the high recall but low precision values. Excluding the last year, the average of the 20 FoS on the remaining 20 timeslots result in Acc = 0.9287, and AUC = 0.9815, and F1 = 0.9287 as shown in Table 4 with the data standardization.

Table 4 Summary of new topic classification results, using standardized data and original data during the training.

Data used	Standardized	Original
Acc	0.9287	0.9240
AUC	0.9815	0.9792
F1	0.9287	0.9243
Precision	0.9522	0.9452
Recall	0.9114	0.9098

The classification performance for the FoS dataset showed slightly higher performance compared to the result based on the journals in the previous research which had an average accuracy of 0.9053 and average AUC score of 0.9809 [18]. Table 4 shows that the same holds even when the original data without standardization are used during the training. This shows that the proposed method is capable of generating highly accurate results with bibliographic datasets built with different criteria and the performance improves when datasets with more focused research interest are used.

The overall performance metrics do not show significant changes in the trend over the years. Accuracy, AUC, and F1 in Figure 1 all share the same pattern over the years, with AUC having a higher average. The sudden drop in *y=2010* can be attributed to the sudden increase in the false positives, having 0.0405 compared to 0.0088 in the previous year. This is reflected in the precision values showing the sharpest change.

The precision and recall intersect around *y=2016* with the changes in the average values of false positive (FP) and false negative results (FN) lower the precision while increasing recall. There are clear differences in the average FP and FN values before and after *y=2016* as shown in Table 5. True positives (TP) increase along with the increase in FP, indicating that the trained model classifies more topics as new in recent years. This average result over 20 diverse FoS topics suggests a possible shift in the overall topic co-occurrence patterns in a specific year, where the neighborhoods of existing topics become more structurally similar to those of new topics over time.

Table 5 Changes in the average TP, FP, FN, precision, and recall of the classification results before and after *y=2016* .

Year	TP	FP	FN	Precision	Recall
< 2016	0.4514	0.0222	**0.0486**	0.9588	0.9028
2016	0.4740	0.0395	0.0260	0.9371	0.9480
> 2016	0.4725	**0.0487**	0.0275	0.9221	0.9450

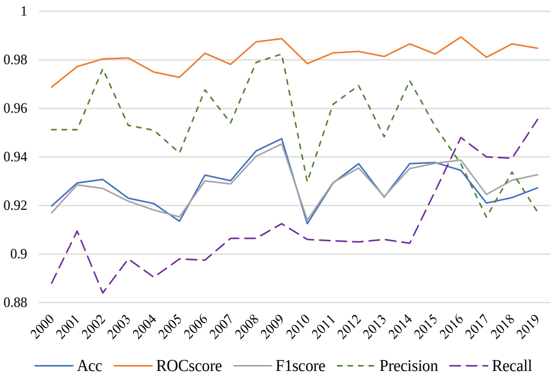

Fig. 1. Binary classification accuracy of *logistic regression* with *y=[2000, 2019]* over 20 FoS datasets.

Experimenting on a different number of features showed that four feature selection functions are statistically similar. ANOVA test was run on the Acc, AUC, precision, recall, and F1 of the classification results for classifications done with *f=1,…,14* using four functions. All 70 ANOVA tests resulted in p=value > 0.9, indicating the differences between the four functions are statistically nonexistent. The result from the *mutual_info_classif* function is used for further analysis.

Figure 2 shows that the number of features used during the training improves the classification performance by a small margin while providing F1 over 0.91 using only one feature. The most significant features are Mean PageRank and Node Count, which were selected for 49% and 50.5% of the 400 classification runs in the experiment. These two features were selected for runs with *f >1*, as well.

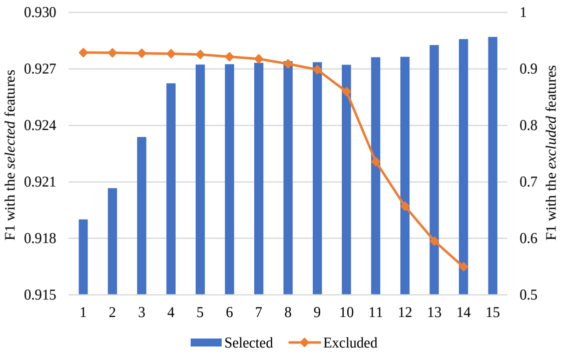

Fig. 2. Changes in the F1 of the classification results using *mutual_info_classif* as the scoring function with *f=[1,…,15]* in the x-axis, with the results of classifications using the *excluded* features shown in the second y-axis.

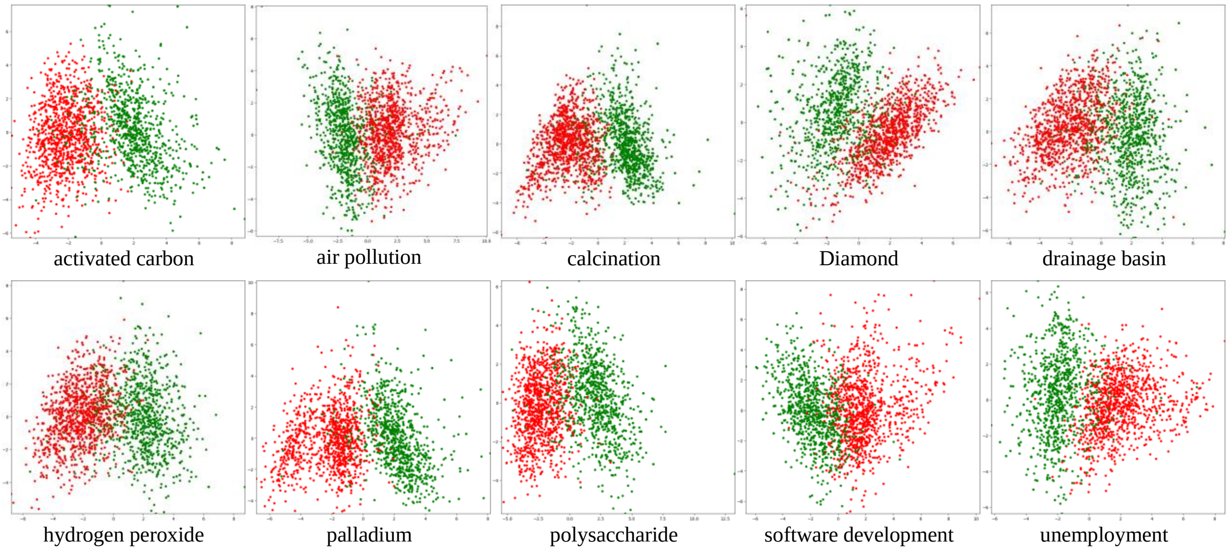

Fig. 3. Visualizations of 2-dimensional PCA results for ten of 20 FoS topics with *y*=2010 and *f*=15. x-axis and y-axis respectively show the first and second features with color-coded labels (green = true, red = false).

The classification results also showed that the result is not dependent on the features. F1 remained at 0.9289 with using only one feature during the training, and the F1 only reaches below 0.9 when 9 out of 15 most significant features were *excluded* during the training. This indicates that the majority of the topic subgraph features are closely correlated to the emergence of a new topic among them, and significant dimension reduction can be done without performance loss.

The PCA results also indicate the possibility of dimension reduction; with 2-dimensional PCA on all 15 features, the first component was able to explain 49.25% of the result while 27.19% were explained by the second component. PCA results of all 20 FoS topics showed more horizontal separations with the first component as the x-axis, with ten randomly selected topics from various fields shown in Figure 3. Clusters of binary labels can be seen in all ten scatterplots. 23.56% of the result remains unexplained by either component, which is likely due to the inclusion of the features with weaker classification strengths. This is shown by the PCA results in Table 6 with feature selections, where lower *f* result in more variance explanations.

Table 6 Ratio of variance explained by 2-dimensional PCA with different *f* selected using *mutual_info_classif*.

f	5	10	15
Explained by 1st component	0.8696	0.6613	0.4925
Explained by 2^{nd} component	0.0742	0.2104	0.2719
Variance Unexplained	0.0562	0.1283	0.2356

5.2. Classifying new topics with incremental learning

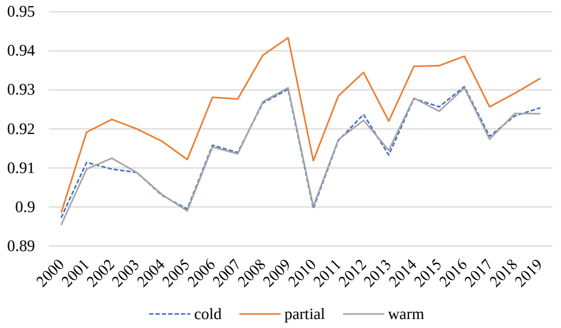

Fig. 4. F1 of classification results for *within-domain* incremental learning over the years.

Figure 4 shows that one of the *within-domain* incremental learning resulted in consistently better results compared to the baseline *cold* approach with the LR algorithm, where the model is re-initialized each year. The *partial* approach resulted in an average of 0.0101 higher F1, showing that there is a temporal consistency over the topic networks for new topic identification. The performance gain increases rapidly during the first 2 years of incremental learning from 0.0078 in *y=2001* to 0.0127 in *y=2002*, and an average of 0.0117 differences was observed until 2015 before being reduced to 0.0072 on average afterward. The performance increase validates the incremental learning over time within a single dataset, while the degree of improvement can vary over time.

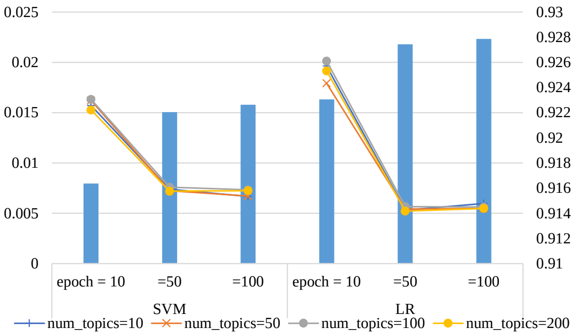

Fig. 5. Changes in the F1 between *partial* and *cold* for different combinations of *epochs* and *num_topics* with the averaged F1 for *partial* as the bar graph in the second axis.

The *warm* approach showed very similar results to the *cold* approach, on the other hand. No apparent performance increase can be attributed to the evolving nature of the topic networks; the connections between predefined topic subsets change every year. The initial training results were overwritten when the ML model is re-trained with such datasets with major shifts, losing any previous training in the process. Using SVM instead of LR resulted in the same outcome, with *partial* with 0.0114 higher F1 and *warm* showing similar values to the baseline, showing statistically insignificant differences for other metrics as well as shown in Table 7. The *warm* approach is statistically identical to the non-incremental learning and hence was removed from further analysis.

Analysis of the different combinations of *epochs* and *num_topics* in Eq. (5) showed the incremental learning can be done with sample sizes smaller than 200. *num_topics* as low as 10 resulted in similar performance improvements with 5 true and 5 false data rows with both classification algorithms as shown by Figure 5, indicating the method can be used even with knowledge domains with sparse topic correlations. The differences were more pronounced with smaller *epochs*, showing higher improvement with lower epoch. This can be attributed to the fact that inadequately trained models have more performance enhancement available to them. More *epochs* resulted in higher absolute performance scores including accuracies and F1, indicating it is still beneficial to train with a larger number of epochs.

Table 7 P-values between *within-domain* incremental learning approach and the baseline.

	LR		SVM	
Pairs	cold/ partial	cold/ warm	cold/ partial	cold/ warm
F1	1.76E-09	**6.92E-01**	2.37E-11	**7.37E-01**
Acc	2.76E-08	**6.84E-01**	5.98E-09	**7.11E-01**
Precision	1.80E-05	**7.43E-01**	7.46E-10	**7.44E-01**
Recall	3.02E-02	**7.09E-01**	9.56E-03	**6.50E-01**

Different FoS datasets resulted in different incremental learning performances. Figure 6 and Figure 7 show the relative F1 improvement of *partial* approach using *epochs=100* and *num_rows=200*, each reaching p = 0.000 for statistical significance. F1 and improvements are reversely correlated, showing moderate to weak correlation with coefficient *corr = -0.5848* for LR and *corr = -0.2758* for SVM. This is in sync with the higher performance gains with lower *epochs*; more improvements are made when possible. While SVM resulted in a higher average improvement of 0.0072 over LR's 0.0055, two of the keywords *cadmium* and *air_polution* showed negative results. LR showed a more consistent performance improvement for all FoS datasets, making it a more generalizable one compared to more dataset sensitive SVM. Consistent improvement for 20 datasets spanning across 14 domains ranging from *business*, *chemistry*, *law* to *medicine* indicates the sequential incremental learning can be done on any field of research to improve new topic identifications.

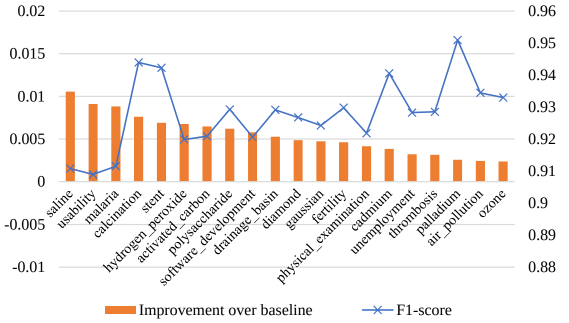

Fig. 6. F1 improvements over the *cold* baseline for individual FoS datasets trained using LR, with F1 in the second axis.

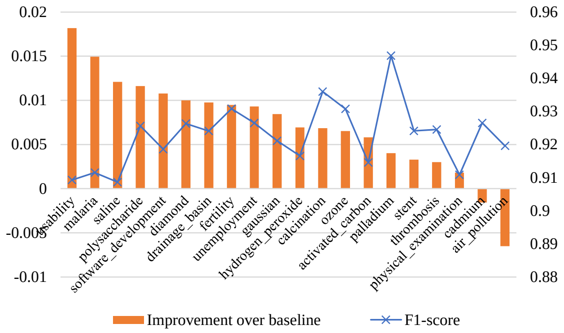

Fig. 7. F1 Improvements over the *cold* baseline for individual FoS datasets trained using SVM, with F1 in the second axis.

The *between-domain* showed that incremental learning can be done over different topics as well. The F1 differences in Table 8 show the performance gain from *between-domain* learning is smaller than that of *within-domain* learning, with one negative value for 48 of the experiment iterations. The performance improvement is also not as statistically significant because of the larger variance in F1 between 380 domain pairs. The t-test between the domain pairs F1 showed an average p-value of 0.6623 for all experiment iterations with baseline *cold* approach (*epochs=[10,10,500]* with *alg=[SVM, LR]*), indicating that there is no inherent difference between the domain pairs. The *partial* approach showed significant differences between the domain pairs with a lower number of epochs, reaching an average p-value of $2.1481e10^{-5}$ using SVM and $6.1439e10^{-9}$ using LR each with *epochs=10*.

The statistical significance diminished with larger *epochs*, with p = 0.0076 for SVM and 0.0523 for LR with *epochs=50* to p > 0.1 for both with *epochs=100*. Such changes in the p-values indicate that the incremental learning over different knowledge domains is harder than the incremental learning done within a single domain; the common knowledge *between-domain* can be acquired with less training compared to the more detailed underlying knowledge *within-domain*. This is supported by the observation that there are no significant differences between the incremental learning done over the domain pairs sharing the same MainType and the ones which do not. The common knowledge captured by the *partial* approach is the basic knowledge common to different domains.

Table 8 Differences in F1 of *between-domain i*ncremental learning approach and the baseline.

Alg	epochs	Year			
		2000	2005	2010	2015
Within the same MainType					
SVM	10	0.0018	0.0033	0.0022	0.0006
	50	0.0023	0.0026	0.0012	0.0018
	100	0.0054	0.0037	0.0035	0.0039
LR	10	0.0047	0.0041	0.0016	0.0028
	50	0.0037	0.0030	0.0028	0.0020
	100	0.0058	0.0052	0.0046	0.0025
Between different MainTypes					
SVM	10	0.0055	0.0003	0.0017	0.0028
	50	0.0029	0.0013	0.0000	0.0027
	100	0.0059	0.0027	0.0029	0.0057
LR	10	0.0029	0.0105	0.0011	-*0.0006*
	50	0.0040	0.0029	0.0038	0.0012
	100	0.0020	0.0045	0.0035	0.0031

6. Conclusion

Topic models derived from processing unstructured documents can capture the number of topics shared throughout a given document collection and can be used to detect and track changes in such topics over time. The text-based approaches however have an innate limitation of requiring the textual data for modeling topics, inhibiting the effective prediction of topic evolutions where such data are nonexistent. The network-based topic emergence identification is an alternative approach utilizing the network structure to model topics, validating the assumption that the new topics can be distinguished by the structural properties of their neighborhoods in the past with classification accuracy up to 0.9.

Binary classification on 20 FoS showed that the proposed method can be applied to bibliographic datasets representing a specific subset of the knowledge domains. The proposed method performed better on topic-specific publications compared to the publications with varying topics of interests. Series of feature selections showed that the proposed method retained F1 over 0.9 with only 6 features; the majority of the 15 topic subgraph features were found to be closely correlated to the emergence of a new topic within them. Analysis of the temporal changes in the classification results showed an underlying topic co-occurrence pattern across diverse research domains; the neighborhoods of existing topics become more structurally similar to those of new topics in more recent years.

Incremental learning is shown to positively affect the results of the proposed method. Consistent performance improvements were observed for incremental learning within each of the 20 FoS over time, showing the method can adapt to various knowledge domains, such as *business*, *chemistry*, *law*, and *medicine*. Iterations of the experiment also revealed that the proposed method can be used even with knowledge domains with sparse topic correlations, retaining similar performance and performance improvements with 10 data instances. The knowledge between different datasets was also found to be transferable with incremental learning between different datasets, albeit with a smaller degree. The common knowledge spanning across different research domains was captured in the early stages of the training, resulting in significant performance improvements only with a smaller number of epochs run during the training.

Future work will include the validation of the method's generalizability with incremental learning results. The shifts in the structural patterns over time can be captured to add explainability to the results, and underlying common structural properties of new topics' neighborhoods will be identified to be incorporated into the prospective new topic prediction, along with the feature selection results. A set of approaches will be made to generate likely neighborhood candidates for the new topic in the future, including community detections and deep neural network optimizations conscious of the properties correlated to the new topic prediction.

References

[1] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, Topic Detection and Tracking Pilot Study Final Report, (1998). doi:10.1184/R1/6626252.v1.

[2] C. Balili, A. Segev, and U. Lee, Tracking and predicting the evolution of research topics in scientific literature, in: 2017 IEEE International Conference on Big Data (Big Data), 2017: pp. 1694–1697. doi:10.1109/BigData.2017.8258108.

[3] C. Battistella, The organisation of Corporate Foresight: A multiple case study in the telecommunication industry, *Technological Forecasting and Social Change*. 87 (2014) 60–79. doi:10.1016/j.techfore.2013.10.022.

[4] D.M. Blei, and J.D. Lafferty, Dynamic Topic Models, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, NY, USA, 2006: pp. 113–120. doi:10.1145/1143844.1143859.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022.

[6] A. Bongers, and J.L. Torres, Measuring technological trends: A comparison between U.S. and U.S.S.R./Russian jet fighter aircraft, *Technological Forecasting and Social Change*. 87 (2014) 125–134. doi:10.1016/j.techfore.2013.12.007.

[7] B. Chen, S. Tsutsui, Y. Ding, and F. Ma, Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval, *Journal of Informetrics*. 11 (2017) 1175–1189. doi:10.1016/j.joi.2017.10.003.

[8] C. Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for Information Science and Technology*. 57 (2006) 359–377. doi:10.1002/asi.20317.

[9] L. Dietz, S. Bickel, and T. Scheffer, Unsupervised Prediction of Citation Influences, in: Proceedings of the 24th International Conference on Machine Learning, ACM, New York, NY, USA, 2007: pp. 233–240. doi:10.1145/1273496.1273526.

[10] J.G. Fiscus, and G.R. Doddington, Topic Detection and Tracking Evaluation Overview, in: Topic Detection and Tracking, Springer, Boston, MA, 2002: pp. 17–31. doi:10.1007/978-1-4615-0933-2_2.

[11] B.C. Gencosman, H.C. Ozmutlu, and S. Ozmutlu, Character n-gram application for automatic new topic identification, *Information Processing & Management*. 50 (2014) 821–856. doi:10.1016/j.ipm.2014.06.005.

[12] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou, Topic Evolution in a Stream of Documents, in: Proceedings of the 2009 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2009: pp. 859–870. doi:10.1137/1.9781611972795.74.

[13] Z. Guo, Z.M. Zhang, S. Zhu, Y. Chi, and Y. Gong, A Two-Level Topic Model towards Knowledge Discovery from Citation Networks, *IEEE Transactions on Knowledge and Data Engineering*. 26 (2014) 780–794. doi:10.1109/TKDE.2013.56.

[14] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, Detecting Topic Evolution in Scientific Literature: How Can Citations Help?, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2009: pp. 957–966. doi:10.1145/1645953.1646076.

[15] S.E. Hug, M. Ochsner, and M.P. Brändle, Citation Analysis with Microsoft Academic, *Scientometrics*. 111 (2017) 371–378. doi:10.1007/s11192-017-2247-8.

[16] Y. Jo, J.E. Hopcroft, and 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, ACM, New York, NY, USA, 2011: pp. 257–266. doi:10.1145/1963405.1963444.

[17] S. Jung, T.M. Lai, and A. Segev, Analyzing Future Nodes in a Knowledge Network, in: 2016 IEEE International Congress on Big Data (BigData Congress), 2016: pp. 357–360. doi:10.1109/BigDataCongress.2016.57.

[18] S. Jung, R. Datta, and A. Segev, Identification and Prediction of Emerging Topics through Their Relationships to Existing Topics, in: 2020 IEEE International Conference on Big Data (Big Data), 2020.

[19] S. Jung, and A. Segev, Analyzing future communities in growing citation networks, in: Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2013) International Workshop on Mining Unstructured Big Data Using Natural Language Processing, ACM, New York, NY, USA, 2013: pp. 15–22. doi:10.1145/2513549.2513553.

[20] S. Jung, and A. Segev, Analyzing future communities in growing citation networks, *Knowledge-Based Systems*. 69 (2014) 34–44. doi:10.1016/j.knosys.2014.04.036.

[21] S. Jung, and W.C. Yoon, An alternative topic model based on Common Interest Authors for topic evolution analysis, *Journal of Informetrics*. 14 (2020) 101040. doi:10.1016/j.joi.2020.101040.

[22] L. Kay, N. Newman, J. Youtie, A.L. Porter, and I. Rafols, Patent overlay mapping: Visualizing technological distance, *J Assn Inf Sci Tec*. 65 (2014) 2432–2443. doi:10.1002/asi.23146.

[23] M. Li, and Y. Chu, Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis, *Journal of Information Science*. 43 (2017) 725–741. doi:10.1177/0165551516661914.

[24] Q. Mei, and C. Zhai, Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, New York, NY, USA, 2005: pp. 198–207. doi:10.1145/1081870.1081895.

[25] N.C. Newman, A.L. Porter, D. Newman, C.C. Trumbach, and S.D. Bolan, Comparing methods to extract technical content for technological intelligence, *Journal of Engineering and Technology Management*. 32 (2014) 97–109. doi:10.1016/j.jengtecman.2013.09.001.

[26] H.C. Ozmutlu, and F. Çavdur, Application of automatic topic identification on Excite Web search engine data logs, *Information Processing & Management*. 41 (2005) 1243–1262. doi:10.1016/j.ipm.2004.04.018.

[27] S. Ozmutlu, Automatic new topic identification using multiple linear regression, *Information Processing & Management*. 42 (2006) 934–950. doi:10.1016/j.ipm.2005.10.002.

[28] A.L. Porter, and M.J. Detampel, Technology opportunities analysis, *Technological Forecasting and Social Change*. 49 (1995) 237–255. doi:10.1016/0040-1625(95)00022-3.

[29] A. Segev, C. Jung, and S. Jung, Analysis of Technology Trends Based on Big Data, in: 2013 IEEE International Congress on Big Data (BigData Congress), 2013: pp. 419–420. doi:10.1109/BigData.Congress.2013.65.

[30] A. Segev, S. Jung, and S. Choi, Analysis of Technology Trends Based on Diverse Data Sources, *IEEE Transactions on Services Computing*. 2015 Vol.8 (2015) 903–915. doi:10.1109/TSC.2014.2338855.

[31] Z. Shen, H. Ma, and K. Wang, A Web-scale system for scientific knowledge exploration, *ArXiv:1805.12216 [Cs]*. (2018). http://arxiv.org/abs/1805.12216 (accessed June 24, 2020).

[32] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. (Paul) Hsu, and K. Wang, An Overview of Microsoft Academic Service (MAS) and Applications, in: Proceedings of the 24th International Conference on World Wide Web, Association for Computing Machinery, Florence, Italy, 2015: pp. 243–246. doi:10.1145/2740908.2742839.

[33] M. Steyvers, and T. Griffiths, Probabilistic topic models, in: Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2007: pp. 427–448.

[34] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, and R. Rogahn, A Review of Microsoft Academic Services for Science of Science Studies, *Front. Big Data*. 2 (2019). doi:10.3389/fdata.2019.00045.

[35] J. Zhang, Z. Ghahramani, and Y. Yang, A probabilistic model for online document clustering with application to novelty detection, in: Proceedings of the 17th International Conference on Neural Information Processing Systems, MIT Press, Vancouver, British Columbia, Canada, 2004: pp. 1617–1624.

[36] C. Zhu, R.H. Byrd, P. Lu, and J. Nocedal, Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, *ACM Trans. Math. Softw.* 23 (1997) 550–560. doi:10.1145/279232.279236.