Tracking and Predicting the Evolution of Research Topics in Scientific Literature

Christine Balili *, Aviv Segev† and Uichin Lee*
*Graduate School of Knowledge Service Engineering
KAIST, Daejeon, South Korea
Email: ccbalili@kaist.ac.kr, uclee@kaist.edu
†CSAIL, MIT, Cambridge, MA, USA
Email: aviv@csail.mit.edu

***Abstract*—The exponential rise in the volume of publications and the prevalence of multidisciplinary practice in scientific domains has made it increasingly difficult to keep track of changes in research trends. In this paper, we propose a framework for determining persistent and emerging research topics in scientific literature. The topics were represented as non-overlapping communities of keywords in a dynamic co-occurrence network derived from 21 million articles in PubMed that were published from 1980 to 2016. We detected a set of communities for each snapshot of the network and traced their instances in consecutive periods using a similarity threshold. Our approach provides a retrospective analysis of changes in research topics: their formation, growth, shrinkage, survival, merging, splitting, and dissolution. We also show that a feature set comprising of 43 temporal and structural attributes from these keyword communities can be used to predict their evolution. In particular, we found that the frequency of co-occurrences and the appearance of new keywords within the community are highly predictive of its persistence or dissolution in the next five years.**

***Keywords*-Community Detection, Community Evolution, PubMed, Dynamic Networks**

I. Introduction

It has been estimated that the amount of scientific literature increases by 8-9% annually [1]. At this rate, the global scientific output doubles every nine years. This unprecedented rise in research productivity implies progress in science, but as a consequence, we must now deal with information overload. Knowledge areas are also becoming more interdisciplinary, so convergent and specialized niches have emerged in recent years. The capacity to recognize patterns and trends in science at an aggregate level is important because funding agencies, academic institutions, and individual researchers can use these insights to develop better strategies to make advancements in the field [2]. However, given the volume, velocity, and variety of scientific publications, identifying promising directions and shifting interests in research has become a highly challenging task.

The evolution of scientific knowledge has been previously studied in terms of communities in co-authorship and citation networks [3]–[5]. Communities are groups of nodes that are densely linked to each other, but are sparsely connected to the rest of the network [6]. In dynamic networks such as those mentioned, the development of persistent communities can be tracked over time. More importantly, these communities have structural and temporal features that can be used to forecast their growth and lifespan and specific events like merging and splitting [7]–[11].

Communities in co-authorship networks are authors that closely collaborate amongst themselves. Thus, changes in their structure characterize social events like project partnerships and advisor-advisee relationships more than topic evolution. Meanwhile, the communities in citation networks correspond to a group of papers assumed to be related to the same topic that has yet to be identified. There are previous works that labeled citation communities based on the most frequent keywords appearing in their papers [12]. However, identifying communities in this manner offers limited insights into the relationships that exist between domain-related concepts. To address these gaps, our work represents research topics as communities of keywords in a dynamic co-occurrence network. We model topic behavior in scientific literature through the evolution of these communities across time: their growth, shrinkage, survival, merging, splitting, and dissolution.

Our investigation uses data from PubMed[1] which is currently the largest repository of biomedical articles from MEDLINE, life science journals, and online books. The proposed approach represents topics as communities in a dynamic network with keywords as nodes and their co-occurrences as weighted edges. The first contribution of this paper is a framework that can historically track the development of research trends through changes in the centrality of nodes and membership size in persistent communities. Second, our method identifies a set of structural and temporal features from communities that can predict how they will evolve within five years.

II. Evolution of Research Topics

We first constructed a dynamic keyword co-occurrence network from 21.18 million articles in PubMed that were published from 1980 to 2016. These keywords come from MeSH (Medical Subject Headings): a controlled vocabulary for indexing articles in PubMed. Since these terms are provided as structured metadata for each article, no further text preprocessing was executed. For our analysis, we only focused on co-occurrences wherein both keywords were tagged as major subjects in more than five articles within a 5-year period.

The dynamic co-occurrence network consists of an ordered set of graphs $\{G_1, G_2, \ldots, G_n\}$, where $G_i = (V_i, E_i)$ is an undirected graph which contains the set of keywords V_i and the set of co-occurrences E_i in period i. We refer to a 5-year window as a **period** or **snapshot**. The edge $(a, b) \in E_i$ has a weight w_i that equals the number of

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

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

articles where keywords *a* and *b* appeared as main subjects in a snapshot. As an example, the edge connecting the nodes *Drug Resistance* and *Malaria, Falciparum* has weight $w = 181$ accounting for the frequency of their co-occurrence from 2010 to 2014. In each graph G_i, we determined n_i disjoint communities $\{C_i^1, C_i^2, \ldots, C_i^{n_i}\}$ that correspond to existing research topics in timestamp *i*. A community C_i^j is also defined as a subgraph (V_i^j, E_i^j). These notations were adopted from [11].

A. Tracking the Evolution of Research Topics

1) Community Detection: Since communities in the network under investigation are not explicitly defined, we employed a detection algorithm that can infer community structure in an unsupervised manner. For this task, we chose **Infomap** [5] based on the criteria proposed by Yang et al. [6]. This algorithm deduces community structure through random walks that track how information flows within the network. Infomap was found suitable for networks with $|V_i| \geq 1000$ and approximate mixing parameter $\mu < 0.50$ [6]. The latter value is the ratio of edges between an initial set of communities determined using Louvain [13].

2) Community Matching and Event Detection: We matched communities found in two consecutive periods based on a similarity threshold θ. The similarity of two communities C_i^j and C_{i+1}^k is defined by Hopcroft et al. [4] as:

$$Sim(C_i^j, C_{i+1}^k) = min\left(\frac{|C_i^j \cap C_{i+1}^k|}{|C_i^j|}, \frac{|C_i^j \cap C_{i+1}^k|}{|C_{i+1}^k|}\right) \geq \theta \qquad (1)$$

A community C_i^j at a given snapshot matches another community C_{i+1}^k in the next snapshot if their similarity is greater than or equal to $\theta = 0.25$. If a community has a match in the subsequent timestep, then we identify it as **persistent**. Otherwise, we say that it has **dissolved**. If a community has no match in the previous period, then we consider it as a **new community** and more specifically, an **emergent** research topic. Persistent communities can be further categorized based on their fluctuation which is defined in [9] as:

$$fluctuation(C_i^j, C_{i+1}^k) = \frac{n_{i+1}^j}{n_i^k} - 1 \qquad (2)$$

where C_i^j and C_{i+1}^k are matching communities with n_i^k and n_{i+1}^j as their number of vertices (or keywords) respectively. We take the event definitions from [9] which are as follows:

Definition 1: A community C_i^j is labeled **survive** if it persists in the next timestep *i+1* and its fluctuation falls within the range of $[-\phi, \phi]$. Note that we set ϕ = 0.10.

Definition 2: A community C_i^j is labeled **growth** if it persists in the next timestep *i+1* and $fluctuation(C_i^j, C_{i+1}^k) > \phi$.

Definition 3: A community C_i^j is labeled **shrink** if it persists in the next timestep *i+1* and $fluctuation(C_i^j, C_{i+1}^k) < -\phi$.

Definition 4: A community C_i^j is labeled **split** if it matches two or more communities in the succeeding timestep.

Table I: Community Structural Features

No.	Feature	Description
1	Size	Number of member nodes
2	Internal Edges	Ratio of the number of internal edges to the total number of edges
3	External Edges	Ratio of the number of external edges to the total number of edges
4	Cohesion	Ratio of the number of internal edges to the number of external edges
5	Weight Inside	Ratio of the total weight of internal edges to the total weight of all edges
6	Weight Outside	Ratio of the total weight of external edges to the total weight of all edges
7	Triangles	Number of triangles in the subgraph (V_i^j, E_i^j)
8-11	Weight Statistics	Mean and standard deviation of the weights w of internal and external edges
12-29	Node Centrality	Mean, standard deviation, and quartiles of the pagerank, closeness, and degrees of the member nodes [14]
30-33	Link Prediction	Mean adamic adar, jaccard coefficient, resource allocation, and preferential attachment scores calculated from a 10% sample of all non-existing edges in the community [14]
34-35	Subgraph Features	Density and average clustering coefficient of the sugraph (V_i^j, E_i^j) [11]

Table II: Community Temporal Features

No.	Feature	Description
36	Size Change	Equivalent to fluctuation; measures the percentage of change in community size
37	Edge Survival	Ratio of the number of edges retained from the previous timestep to the current number of edges in the community
38	Edge Mortality	Ratio of the number edges that disappeared in the current snapshot to the previous number of edges in the community
39	Retention	Ratio of the number of nodes that remain present in the current instance of the community to its size in the previous period
40	New Nodes	Ratio of the number of new nodes to the community's current size
41	New In-edges	Ratio of the number of new internal edges to the total number of new edges in C_i^j
42	New Out-edges	Ratio of the number of new external edges to the total number of new edges in C_i^j
43	Previous Event	Event that occurred to the community in the previous period: grow, shrink, surivive, split, or merge

Definition 5: Two or more communities at timestep *i* are labeled **merge** if they match the same community in the succeeding timestep.

B. Predicting the Evolution of Research Topics

1) Feature Extraction: The task of predicting the evolution of a research topic can be considered as a supervised classification problem. Communities at timestep *i* are labeled based on the events that occurred to them in the subsequent period. We extracted temporal and structural properties from communities to classify them into two classes: **persist** or **dissolve**.The structural features that are enumerated in Table I are intended to measure the strength of connections among community components in the present period. Table II lists temporal features which reflect changes in community properties using its former instance as a point of reference.

2) Classification: Based on the derived features, we used a supervised classification model to predict whether a community will **persist** or **dissolve** in the next five years. The experiments were performed using Support Vector Machines (SVM), Gradient Boosted Trees (XGBoost), and Logistic Regression. The most important features were distinguished through a tree-based estimator. We performed stratified 4-fold cross validation on 391 communities composed of at least 10 keywords. There were 197 samples for persist events and the remaining 194 were for dissolved cases. These

communities were taken from five snapshots starting from 1985 to 2009. Due to the dependence of temporal features on the previous instance of a community, the newly formed ones were excluded in the prediction task. Finally, we examined how the major research topics that were identified in the period 2010-2014 will change in 2015-2019 in light of the resulting classification model.

III. Results and Discussion

A. Retrospective Analysis of Research Themes

1) Persistent Research Topics in Biomedicine (1980-2014): Community detection at each snapshot of the co-occurrence network allowed us to recognize major research topics per period. The communities were identifed based on their nodes that had the highest pageranks. We consider the number of keywords n that comprise a topic as a measure of its extent. Based on this assumption, the five biggest research areas in biomedicine as of 2016 are *Cancer Research* ($n = 2049$), *Mental Disorders* ($n = 1808$), *Drug Synthesis* ($n = 838$), *Medicinal Plants* ($n = 810$), and *Neuroscience* ($n = 745$).

We were able to trace the continuity of a research topic for successive timesteps. There were several keyword communities identified at the beginning that persisted throughout the duration of our analysis which spanned a total of 37 years (1980 - 2016). Most of these persistent communities exhibit growth over time. For example, the community representing *Anti-Bacterial Agents* was initially composed of 251 MeSH terms from 1980-1984; its size has increased to 720 keywords in 2010-2014. The aforementioned major research topics consistently follow this trend. Community enlargement can be interpreted as the expansion and specialization of a research area. In terms of publications, this implies that new associations are formed and new concepts are being studied in the field. There are also persistent communities that show continuous shrinkage such as *Occupational Diseases* which started with 118 nodes in 1980-1984 but has reduced to 73 in 2010-2014. This phenomenon can signify the decrease of interest in this topic, and hence, the pool of keywords that have shrunk over time. This could also be explained by the increased co-occurrence of keywords originally belonging to this topic with keywords from another topic resulting to a change in membership from the former to the latter.

2) Emerging Research Topics in Biomedicine: Our framework has also distinguished emerging topics: communities that have not existed previously but are composed of at least 100 keywords in the current period. In 1995-1999, the community represented by *Medical Records Systems*, *Information Systems*, and the *Internet* was established. This aligned with the considerable adoption of computer systems into medical practice and hospital management and widespread use of the internet. From 2015-2016, research on *Electronic Health Records*, *Information Storage and Retrieval*, and *Data Mining* has formed into a major research area in biomedicine ($n = 100$). In addition to detecting

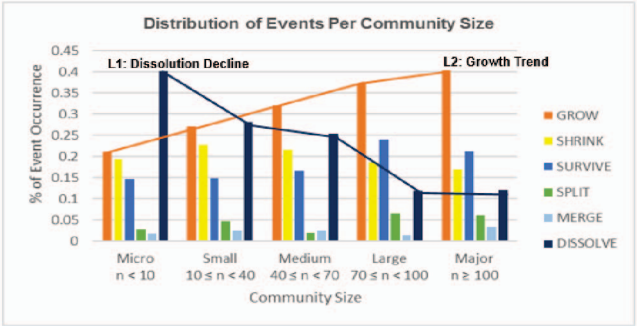

Figure 1: Percentage of the occurrence of events across community sizes. L1 shows that the frequency of the dissolution events tend to decrease in bigger communities. An opposite trend is observed for growth events as depicted in L2.

emerging themes in the domain at large, we also found that shifts in research interests within a topic can be observed by examining the change in the centrality of its member nodes. In Table III, we list the nodes that have the highest pageranks from the topic *Anti-Bacterial Agents* across three periods. Based on this ranking, the concern on *Drug Resistance* became more apparent after 1995 when this node gained higher centrality. Table IV further supports this point by showing the top five nodes with the highest pageranks in the topic *Disease Outbreaks*. From 2010-2014, *Influenza A virus, H1N1 Subtype* became a "hot" subject as a response to an outbreak in 2009 [15]. Although not shown here, the community on *Image Processing* first emerged as major topic in 2000-2004. It is interesting to note that *Pattern Recognition* and *Artificial Intelligence* have garnered high centrality to this area since then.

The percentage of occurrence of each event across community sizes is displayed in Figure 1. The increasing trend in L2 shows that larger communities are more likely to experience persistence and growth. This finding concurs with the phenomenon of preferential attachment ("rich gets richer") in citation networks [12]. The frequency of dissolution events follows the opposite; it decreases as one moves up to bigger size groups as depicted in L1. The merging and splitting of communities are less frequent events because persistent communities tend to stay cohesive. However, in 1990-1994, we did observe the merging of *Coronary Heart*

Table III: Evolution of *Anti-bacterial Agents*

1980-1984	1995-1999
Anti-Bacterial Agents	Anti-Bacterial Agents
Bacterial Infections	Staphylococcal Infections
Bacteria	Penicillins
Neonatal disorder	Bacteria
Cephalosporins	Antibiotics, Antitubercular
Disease Outbreaks	Anti-Infective Agents
Staphylococcal Infections	Genus staphylococcus
Cross Infection	Disease Outbreaks
Sepsis	**Drug Resistance, Microbial**

Table IV: Evolution of *Disease Outbreaks*

2005-2009	2010-2014
Disease Outbreaks	Influenza
Influenze	Disease Outbreaks
Antibodies, Viral	**Influenza A Virus, H1N1 Subtype**
Vaccination	Vaccination
Viral Vaccines	Antibodies, Viral

Disease and *Arterial Occlusive Diseases* which are closely related topics. Merging events can signal the convergence of closely related or more interestingly, separate research areas over time. On the other hand, the splitting of a community denotes the decomposition of a topic into more specialized areas. An example of this would be the research on *Reproductive Health*. In 2010-2014, this field has split into two communities: *Contraceptive Methods* and *In-vitro Fertilization*.

B. Community Evolution Prediction

The highest accuracy and F1 score for predicting whether topic communities will **persist** or **dissolve** using the proposed set of features was achieved by Gradient Boosted Trees at 69.5% and 69.48% respectively, but this is only slightly better than that of SVM and Logistic Regression. An examination of the confusion matrix indicates that small and medium communities are more frequently misclassified than large and major communities. We can attribute this confusion to the almost equal frequency of **growth** and **dissolve** events in the former community sizes as shown in Figure 1. In general, it appears that the evolution of smaller research areas are harder to predict under our framework.

Based on a tree-based estimator, the five most informative features for this classification problem consists of the structural attributes (Features 2-6): weight inside, weight outside, internal edges, external edges, and cohesion. These are followed by new nodes (40), new in-edges (41), and previous event (43) which are temporal features. Our classifier predicts that most major research themes existing in 2010-2014 will persist in 2015-2019 except for *Image Processing*. This prediction does not imply that this topic will dissolve entirely. A more sensible explanation could be that this area will decompose into smaller and more specialized topics as artificial intelligence becomes more integrated into biomedical research.

IV. Conclusion

In this study, we represented research topics as communities of keywords in a dynamic co-occurrence network. Based on a large amount of scientific publications, our approach has provided a retrospective and predictive analysis of how these topics will develop over time. The progress of research in an area is reflected by its persistence and growth as a community of keywords. Furthermore, we have shown that the evolution of the research landscape is not only apparent in terms of community events, but also in the change of centrality among member nodes or keywords. This framework draws the big picture from millions of published work in a domain, and thus provide insights into shifting research trends in science. We believe that this approach can be further applied to other scholarly archives with an existing ontology of technical terms. As future work, we intend to explore how this paradigm can be used for information visualization. We also aim to build a burst detection and prediction model based on cliques in the dynamic network for hypothesis recommendation purposes.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

References

[1] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis," *CoRR*, vol. abs/1402.4578, 2014. [Online]. Available: http://arxiv.org/abs/1402.4578

[2] J. A. Evans and J. G. Foster, "Metaknowledge," *Science*, vol. 331, no. 6018, pp. 721–725, 2011.

[3] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, no. 11, pp. 758 – 775, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166497208000436

[4] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5249–5253, 2004.

[5] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008. [Online]. Available: http://www.pnas.org/content/105/4/1118.abstract

[6] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Scientific Reports*, vol. 6, 2016.

[7] G. Palla, A.-L. Barabasi, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007. [Online]. Available: http://dx.doi.org/10.1038/nature05670

[8] M. Goldberg, M. Magdon-Ismail, S. Nambirajan, and J. Thompson, "Tracking and predicting evolution of social communities," in *IEEE Third Inernational Conference on Social Computing*. IEEE, 2011, pp. 780–783.

[9] N. lhan and S. G. Oguducu, "Feature identification for predicting community evolution in dynamic social networks," *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 202 – 218, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197616301117

[10] S. Saganowski, B. Gliwa, P. Bródka, A. Zygmunt, P. Kazienko, and J. Kozlak, "Predicting community evolution in social networks," *CoRR*, vol. abs/1505.01709, 2015. [Online]. Available: http://arxiv.org/abs/1505.01709

[11] M. Takaffoli, R. Rabbany, and O. R. Zaïane, "Community evolution prediction in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014, pp. 9–16.

[12] S. Jung and A. Segev, "Analyzing future communities in growing citation networks," *Knowledge-Based Systems*, vol. 69, pp. 34 – 44, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095070511400166X

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[14] C. C. Aggarwal, "An introduction to social network data analytics," *Social Network Data Analytics*, 2011.

[15] W. H. Organization, "What is the pandemic (h1n1) 2009 virus?" 2009. [Online]. Available: http://www.who.int/csr/disease/swineflu/frequently_asked_questions/about_disease/en/