What Kind of Research Topics Emerged in the Biomedical Domain?: A Perspective from Newly Added Subject Terms in a Thesaurus

Kun Lu [0000-0001-5614-7042]

University of Oklahoma, Norman OK 73019, USA
kunlu@ou.edu

Abstract. Understanding what kind of research topics will emerge in scientific domains has implications for researchers, policy makers, and industry. Existing studies on emerging topics generally approach the problem from the publication perspective, that is, to retrieve a set of publications and identify emerging topics from them. This study aims to offer a different perspective to understand emerging topics. By tracing the newly added subject terms in a thesaurus, this study provides insights into what kind of topics emerged and what did not. This study found that economic, biomedical, or social impact is important for a biomedical topic to emerge in addition to the common features used for identifying emerging topics. More attention was paid to topics that have direct impact on humans than on other species. This aspect is largely neglected in the literature. The findings from this study have implications for designing predictors for emerging research topics in the biomedical domain.

Keywords: Emerging Topics, Thesaurus, Medical Subject Headings, New Subject Terms.

1 Introduction

Research topics in scientific domains evolve progressively. New topics constantly emerge and old topics gradually fade out. However, not all new topics will flourish and sustain popularity over time. Identifying emerging research topics and predicting their future fate has become an interesting research topic itself due to its implications for researchers, policy makers, and industry. There is not yet a unified definition of emerging topics. An emerging topic is often characterized with novelty and rapid growth. Additional attributes of emerging topics, such as coherence, prominent impact, and uncertainty [1], have also been discussed. Different methods for identifying emerging topics have been proposed, such as citation- and co-citation-based approach [2], co-word analysis-based approach [3], growth trend analysis [5], and machine learning-based approach [6]. Different sources have been used to characterize emerging topics, including free-text terms from titles, abstracts, and/or full-text [4], subject terms from controlled vocabularies [3], and latent topical structures [8].

Existing studies on emerging topics generally approach the problem from the perspective of publications, that is, to retrieve a set of publications/patents from one or more databases over a certain period of time (e.g. a ten-year period), and then to identify emerging topics from the dataset. In this study, a different perspective is offered to study emerging topics and their fate from the time new subject terms are added to a thesaurus. Subject terms in a thesaurus are created to describe literature in a field and reflect the knowledge structure. Examining the newly added subject terms provides a more complete story of the development of the topics since their inclusion in the thesaurus. It also shows which terms have not taken off since inclusion. Another advantage of this perspective is that when adding new subject terms, professionals have already assessed the necessity of the addition based on principles of thesaurus management, such as literary warrant, user warrant, and structural warrant et al. [7]. Therefore, there is already a manual screening process on what can become new subject terms.

2 Method

MeSH (Medical Subject Headings) thesaurus is a controlled and hierarchically-organized vocabulary in the biomedical field to index publications in PubMed. It is maintained by the National Library of Medicine and updated regularly. This study selected 89 MeSH terms introduced in 2001 and 2002, and retrospectively examined their popularity after being added to the thesaurus. The popularity is measured by the number of articles indexed with the MeSH terms.

2.1 Data Collection

For the data collection process, the lists of new MeSH terms in 2001 and 2002 were retrieved from ftp://nlmpubs.nlm.nih.gov/online/mesh/1999-2010/newterms/. There are 184 and 847 new terms on the 2001 and 2002 lists, respectively. For each new term, we looked it up in the MeSH Browser (https://meshb.nlm.nih.gov/search) and only kept the term that does not have a “Previously Indexing” term, because if a “Previously Indexing” term exists, it indicates that the MeSH term is an existing concept indexed under a different name, while this study aims to track new concepts. Then, an advanced search was performed in the MeSH Advanced Search Builder (https://www.ncbi.nlm.nih.gov/mesh/advanced) with the options “Restrict to MeSH Major Topic” and “Do not include MeSH terms found below this term in the MeSH hierarchy” checked to search the publications in PubMed that are indexed with the subject term. The first article that was indexed with the MeSH term was identified. If the publication date of the article is more than five years before the subject term was added, the subject term was excluded because this indicates an existing concept. With this, the study selected a total of 89 newly added MeSH terms, 20 from 2001 and 69 from 2002. Searches were then performed in PubMed for articles that were indexed with the MeSH terms by the end of 2017.

Figure 1 summarizes the data collection process in a diagram.

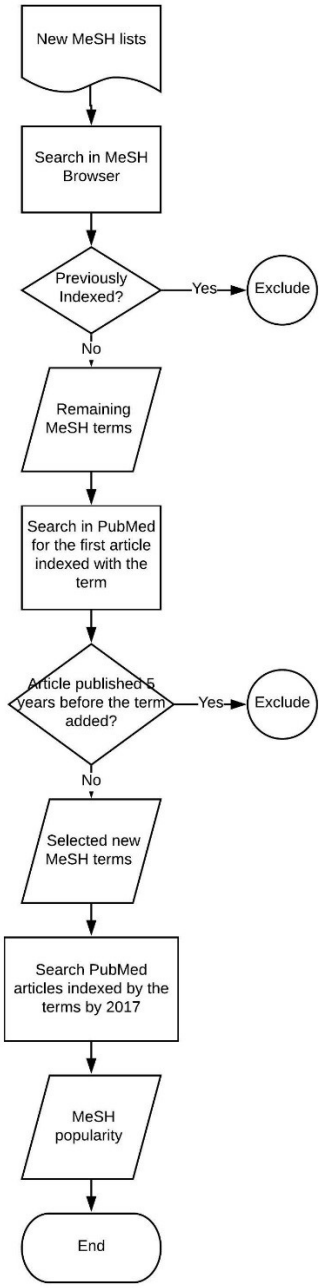

Fig. 1. Data collection process.

3 Results

3.1 Descriptive analysis

The number of articles indexed by the MeSH terms is used as a measure of their popularity. Table 1 lists descriptive statistics for the total number of articles indexed by the MeSH terms and the average number of articles indexed by the MeSH terms per year until 2017. The most popular MeSH term is "Xenograft Model Antitumor Assays". Since its inclusion in 2001, it was assigned to 1391 articles with an average of 77.28 per year (the term was first used for indexing in 2000).

Table 1. Descriptive statistics of MeSH popularity (*N*=89).

	Mean	**Min.**	**1st Quartile**	**Median**	**3rd Quartile**	**Max.**
Total # of articles indexed by the MeSH terms	154.4	2	17	58	161	1391
Average # of articles indexed by the MeSH terms per year	8.337	0.125	1	3.222	8.812	77.278

Table 2. The most popular new MeSH terms in 2001 and 2002 since inclusion.

MeSH term	**Total # of articles indexed**
Xenograft Model Antitumor Assays	1391
Chlamydophila Infections	1162
Food, Genetically Modified	1156
Metapneumovirus	964
Bystander Effect	891

Table 2 and 3 list the most and least popular new MeSH terms from our dataset. Among the popular ones, "Xenograft Model Antitumor Assays" is an anticancer drug screening technique that transplants human tumor tissue into mice or rats, and then assesses the effectiveness of tumor treatment there. The popularity of this term indicates the technique is widely used in anticancer drug screening, and also there is a great interest in developing anticancer drugs. Both "Chlamydophila Infections" and "Metapneumovirus" can cause respiratory diseases, while the former is caused by bacteria and the latter is caused by virus. The popularity of these two terms indicates the interest in treating diseases. "Food, Genetically Modified" is a pretty hot topic with wide public interest in its potential risks. "Bystander Effect" is closely related to drug effect. The popular terms appear to have a commonality that they are all very closely related to our lives and have wide economic, biomedical, or social impact.

Table 3. The least popular new MeSH terms in 2001 and 2002 since inclusion.

MeSH term	Total # of articles indexed
Hyperotreti	2
Astragalus gummifer	2
Primate T-lymphotropic virus 2	2
Ilex guayusa	3
Esociformes	4

Least popular terms seem to be all related to non-human organisms. "Hyperotreti" is a group of invertebrate chordates in the subphylum Craniota according to its scope note. The term has a narrower term on the MeSH hierarchy: Hagfishes (established in 1991). Hagfishes is much more popular with 453 articles by 2017. Without knowing the details of MeSH staff's decision making, it is conjectured that the addition of "Hyperotreti" was based on the structural warrant [9] to account for other kinds of species under the umbrella in the future. According to the NCBI Taxonomy database, the lineage of this line has Hyperotreti; Myxiniformes; Myxinidae (which is Hagfishes). Hagfishes seems to be a more commonly used term by the public. "Astragalus gummifer" is a shrub. Although added in 2002, the first article indexed by the term in PubMed was published in 2014. With only two articles indexed by the term, it is a very unpopular topic. However, "Astragalus gummifer" produces a special gum called Tragacanth that attracts a much higher attention (99 articles) likely due to its use in foods, cosmetics, and pharmaceuticals. The next term "Primate T-lymphotropic virus 2" was introduced in 2002, but the first article indexed by the term was not published until 2005. Looking into the term, it is a species of virus in the family of deltaretrovirus, includes HTLV-2 and STLV-2. One of its narrower term "Human T-lymphotropic virus 2" (HTLV-2) received much more attention (572 articles by 2017), while "Simian T-lymphotropic virus 2" (STLV-2) has no article indexed by it in PubMed. Apparently, there is much greater interest in human-related virus than other species-related. It is conjectured that the term "Primate T-lymphotropic virus 2" was also added based on the structural warrant. "Ilex guayusa" is a plant species introduced in 2002 with first article published in 2016. Its leaves make tea that is generally consumed in Ecuador, Peru, and Colombia [10]. It also has ritual use by Amazonian Jivaro Indians according to its scope note. "Esociformes" is an order of fishes native to North America and Northern Eurasia. It has two narrower terms: "Esocidae" (194 articles) and "Umbridae" (9 articles). Apparently, more interest lies in the "Esocidae" rather than its broader term "Esociformes".

Based on the observation, it appears that new species that do not have direct or immediate influence on humans, such as being related to diseases, drugs, or other benefits, do not receive as much attention. And as a result, they do not become emerging topics. The preliminary findings suggest that social and economic impact is very important for a topic to become emerging in the biomedical domain.

3.2 Different trends

Different groups of new subject terms can be formed qualitatively based on their trends. Some terms emerged and sustained their popularity (group 1), some others emerged but failed to sustain (group 2), and others had not become popular since inclusion (group 3). Fig. 2-4 provide examples for each group. Topics in group 1 attract and sustain wide attention. These are productive topics that continue generating impact. Topics in group 2 emerged but lost popularity. It can be the problem is solved, or is not solvable at the moment. Topics in group 3 have not yet garnered wide attention. It could be due to the problem is not of great interest, or researchers have not realized its importance. In addition to the topic itself, a new subject term may also be added due to structural considerations rather than literary need.

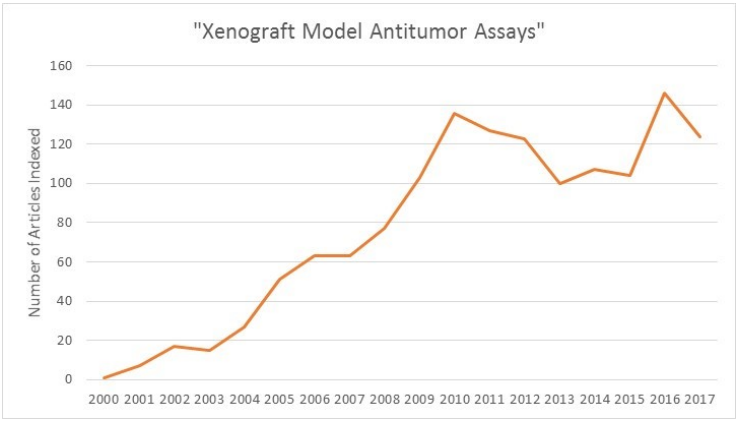

Fig. 2. An example of emerged and sustained topics (group 1).

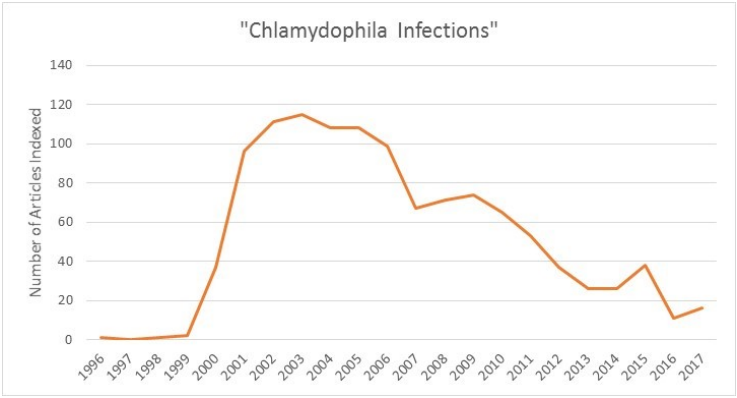

Fig. 3. An example of emerged but not sustained topics (group 2).

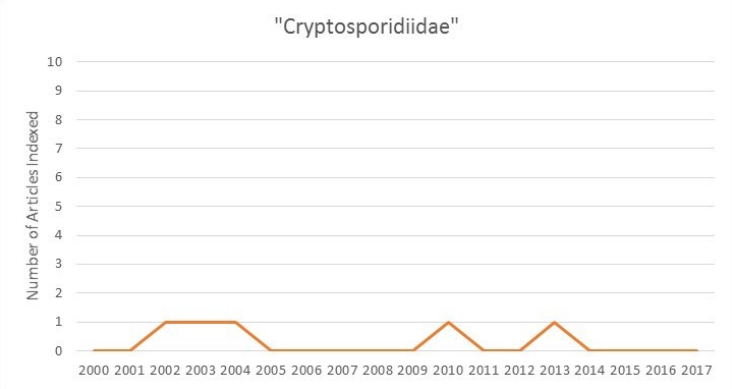

Fig. 4. An example of not yet emerged topics (group 3).

4 Conclusions

This preliminary study retrospectively examined newly added subject terms and their popularity since introduction. It provides insights into the popular and unpopular topics. The findings suggest that economic and social impact is important for a biomedical topic to become popular in addition to the bibliometric, semantic, and structural features existing studies commonly explore to identify emerging topics. Much more attention was paid to topics that have direct impact on humans than on other species. Popular topics tend to have wide economic, biomedical, or social impact. However, current methods of identifying emerging topics generally overlooked this aspect.

Certainly, the findings are limited to the biomedical domain and not ready to generalize to other domains. In addition, the coverage of PubMed has an impact on the results. According to [11], PubMed primarily covers citations in biomedicine and health fields, and related disciplines such as life sciences, behavioral sciences, chemical sciences, and bioengineering. This may be the reason why many popular topics are disease- or drug-related. However, it is still interesting to see that among health-related topics, there are different trends and variations in popularity. In addition, the granularity of domain selection influences emerging topics. This study selected PubMed that represents life sciences and biomedical domain in general. However, some subdomains that PubMed covers, such as biodiversity or evolutionary biology, are smaller than other subdomains, such as biomedicine. New MeSH terms in these smaller subdomains are less likely to be as popular as those in larger subdomains, and thus are less likely to be identified as emerging topics. To understand the details in subdomains, one will need to zoom in and study at a more granular level. It should also be noted that the goal of this study is not to predict emerging topics, but to understand the features of different topics since they are included in the thesaurus. The findings from this study have implications for designing predictors for emerging topics.

5 Acknowledgement

The author acknowledges the assistance of Chloe Summers in collecting the data for this study.

References

1. Rotolo, D., Hicks, D., Martin, B.: What is an emerging technology? Research Policy 44(10), 1827-1843 (2015).
2. Small, H., Boyack, K.W., Klavans, R.: Identifying emerging topics in science and technology. Research Policy 43(8), 1450-1467 (2014).
3. Ohniwa, R.L., Hibino, A., Takeyasu, K.: Trends in research foci in life science fields over the last 30 years monitored by emerging topics. Scientometrics 85(1), 111-127 (2010).
4. Porter, A.L., Garner, J., Carley, S.F., Newman, N.C.: Emerging scoring to identify frontier R&D topics and key players. Technological Forecasting & Social Change 146, 628-643 (2019).
5. Ho, J.C., Saw, E.C., Lu, L.Y.Y., Liu, J.S.: Technological barriers and research trends in fuel cell technologies: A citation network analysis. Technological Forecasting & Social Change 82, 66-79 (2014).
6. Lee, C., Kwon, O., Kim, M., Kwon, D.: Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technological Forecasting & Social Change 127, 291-303 (2018).
7. Barité, M. “Literary warrant”. Knowledge Organization 45(6), 517-536 (2018).
8. Ranaei, S., Suominen, A.: Using machine learning approaches to identify emergence: Case of vehicle related patent data. In: Proceedings of 2017 Portland International Conference on Management of Engineering and Technology (PICMET), pp. 1-8. IEEE, Portland, OR (2017).
9. Svenonius, E.: Design of controlled vocabularies. Encyclopedia of library and information science 45(suppl 10), 82-109 (1989).
10. Wikipedia page, https://en.wikipedia.org/wiki/Ilex_guayusa#Uses, last accessed 2019/8/27.
11. PubMed Overview, https://www.nlm.nih.gov/bsd/pubmed.html, last accessed 2019/11/22.