DE GRUYTER
OPEN

Research Paper

A Micro Perspective of Research Dynamics Through "Citations of Citations" Topic Analysis

Xiaoli Chen[1,2†], Tao Han[1,2]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Citation: Chen, X.L., & Han, T. "A micro perspective of research dynamics through 'citations of citations' topic analysis." *Journal of Data and Information Science*, vol. 5, no. 4, 2020, pp. 19–34. https://doi.org/10.2478/jdis-2020-0034

Received: Feb. 8, 2020
Revised: May 20, 2020
Accepted: Jun. 11, 2020

Abstract

Purpose: Research dynamics have long been a research interest. It is a macro perspective tool for discovering temporal research trends of a certain discipline or subject. A micro perspective of research dynamics, however, concerning a single researcher or a highly cited paper in terms of their citations and "citations of citations" (forward chaining) remains unexplored.

Design/methodology/approach: In this paper, we use a cross-collection topic model to reveal the research dynamics of topic disappearance topic inheritance, and topic innovation in each generation of forward chaining.

Findings: For highly cited work, scientific influence exists in indirect citations. Topic modeling can reveal how long this influence exists in forward chaining, as well as its influence.

Research limitations: This paper measures scientific influence and indirect scientific influence only if the relevant words or phrases are borrowed or used in direct or indirect citations. Paraphrasing or semantically similar concept may be neglected in this research.

Practical implications: This paper demonstrates that a scientific influence exists in indirect citations through its analysis of forward chaining. This can serve as an inspiration on how to adequately evaluate research influence.

Originality: The main contributions of this paper are the following three aspects. First, besides research dynamics of topic inheritance and topic innovation, we model topic disappearance by using a cross-collection topic model. Second, we explore the length and character of the research impact through "citations of citations" content analysis. Finally, we analyze the research dynamics of artificial intelligence researcher Geoffrey Hinton's publications and the topic dynamics of forward chaining.

Keywords Research dynamics; Forward chaining; Topic model; Scientific influence; Citations content analysis

† Corresponding author: Xiaoli Chen (E-mail: chenxl@mail.las.ac.cn).

Research Paper

1 Introduction

Research dynamics and topic evolution are hot topics in Scientometrics. These research works aim to discover the temporal research trends of a certain discipline or subject. Most previous works only focus, however, on the dynamics of inherited topics and the innovative topics of follow-up studies (through direct citations), but disappearing topics of the cited work remain unexplored. Disappearing topics are those that are not inherited (mentioned) by follow-up studies. These topics might be potentially emerging technologies in certain situations or "sleeping beauty" candidates for the future. Revealing the dynamic of disappearing topics is an interesting point. In this paper, we give our solution by using cross-collection topic models (ccTM). In ccTM, common topics and inclusive topics are drawn from collection independent words and collection-dependent words, respectively. In our use case, we draw disappearing topics, inherited topics and innovative topics from cited collection specific words, cited and citation shared words, and citation collection specific words, respectively.

Most of the studies on academic influence try to find an explicit answer to the impact of original work on follow-up research. For research elites or highly cited work, this impact may be long-lasting, not only directly on citations but also indirectly on citations of citations (Hu, Rousseau, & Chen, 2011). In this paper, we use forward chaining to reveal this impact and how long it lasts (as measured by citation generations).

This paper is organized as follows. In Section 2, we introduce related work, especially work on the cross-collection topic model and its variations. We give a detailed explanation of our topic model in Section 3. In Section 4, we use our model to reveal the research dynamics over each generation of forwarded citations stemming from artificial intelligence researcher Geoffrey Hinton.

2 Related work

2.1 Research dynamics method

To perceive or even predict temporal changes to the research landscape, researchers from various domains leverage topic models to revise the research dynamics of their domains. Beykikhoshk et al. (2016) use topic model to analyze the history of the autism spectrum disorder. Doyle and Elkan (2019) use Dirichlet compound multinomial (DCM) distributions to model the phenomenon of "burstiness.". Wu et al. (2014) the study topic evolution of stem cell research literature. De Battisti, De Battisti, Ferrara and Salini (2015) analyze the links between topics and their temporal evolution. Wu et al. (2010) use topic models to

mine the literature in the field of bioinformatics and discover important research topics, quantifying the evolution of these themes and showing their trends. Hall, Jurafsky and Manning (2008) apply topic modeling to analyze historical trends in the field of Computational Linguistics from 1978 to 2006. Yan (2015) uses topic-based approaches to reveal research dynamics, impacts and the dissemination of informetric studies.

2.2 Topic models for cross-collection datasets

There are various topic models since Blei et al. (2003) first introduced Latent Dirichlet Allocation (LDA). Since we want to demystify research dynamics through forward citations analysis, we are especially focused on topic models that can model correlations or differences between datasets. The first such models were Dynamic Topic Models like those of Gerrish and Blei (2010), Iwata et al. (2010), and Xu et al. (2014). These models incorporate time-varying variables in the topic model to reveal the temporal change of topics. The second such models were correlation topic models like that of Chang (2009), Chang and Blei (2010), and Li and Mccallum (2006). Unlike topic models that model two-layer multinomial distributions as a latent Dirichlet allocation they use Pachinko allocation to model correlations between topics, which are assumed to be in latent Dirichlet allocation.

Our work is based on cross-collection topic models (ccTM). These models include ccMix (Zhai, Velivelli, & Yu, 2004), ccLDA (Paul & Girju, 2009, 2010), the Inheritance Topic Model (He et al., 2009), Link-PLSA-LDA (Nallapati & Cohen, 2008; Nallapati et al., 2008), the Differential Topic Model (Chen et al., 2015), RefTM (Shen, 2016), C-LDA and C-HDP (Zhang et al., 2015), cite-LDA and cite-PLSA-LDA (Kataria, Mitra, & Bhatia, 2010), LTAI (Kim, Kim, & Oh, 2017), citation-LDA (Wang, Zhai, & Roth, 2013), content-citation LDA (Zhou, Yu, & Hu, 2017) and the entropy-based topic model (Risch & Krestel, 2018). The most similar work (Dietz, Bickel, & Scheffer, 2007) devises a probabilistic topic model that explains the generation of documents; this model incorporates the aspects of topical innovation and topical inheritance via citations.

To solve our problem of comparing topics of cited collection and citation collection, we borrow the idea of these cross-collection models in which common topics and inclusive topics are drawn from collection-independent words and collection-dependent words, respectively. In our topic model, we draw disappearing topics, inherited topics and innovative topics from cited collection specific words, citation collection independent words and citation collection specific words.

Studies like those of Elgendi (2019), Martínez et al. (2015), and Parker, Allesina, and Lortie (2013) also try to characterize highly cited papers by their citation counts,

views and other features. Based on our previous work (Chen & Han, 2019), we go further behind the citation statistics and analyze citation content by extending the ccTM model to find out what impact highly cited researchers have on follow-up studies, as well as how far their impacts go by measuring the number of influenced generations of forward citations.

3 Methodology

We use cross-collection topic models (ccTM) to solve the problem of topics regarding several collections. In addition to independent document-topic and topic-word distributions of each collection, ccTM also tries to capture the common shared knowledge among collections and the unique knowledge of each collection. This is possible because we can observe collection-dependent words and collection-independent words.

In our study, we have two collections of documents: cited documents and citation documents. We have observations of cited-independent words (disappearing words), shared common words (inheritance words) and citation-independent words (innovative words). Cited collection documents are a combination of disappearing words and inherited words and citation collection documents are a combination of inheritance words and innovative words. In Section 4, we will introduce how we apply ccTM to reveal topic disappearance, topic inheritance, and topic innovation of cited collection and citation collection.

We state our problem as follows: given two collections of scientific articles, one collection is the cited collection, and the other is the citation collection. D is the number in the documents of citation collection. These two collections share K topics globally expressed over V unique words. θ_d is the k dimensional document-topic distribution of the citation document d. μ_d is the k dimensional document-topic distribution of cited document d. σ_d is the k dimensional shared document-topic distribution between cited document d and the citation document d. ϕ is the V dimensional unique topic-word distribution of the citation document d. It is shared between the cited collection and the citation collection.

The ccTM model uses a shared topic-word distribution and a collection-specific topic-word distribution. In our use case, the shared topic-word distribution represents inherited topics and the cited specific topic-word distribution represents disappearing topics. Finally, the citation specific topic-word distribution represents innovative topics. We use these distributions to generate cited specific document-words, shared cited document-words and citation-specific document-words based on word observations. These three document-words should also combine to converge at the cited document-word observation and citation document-word observation at the same time.

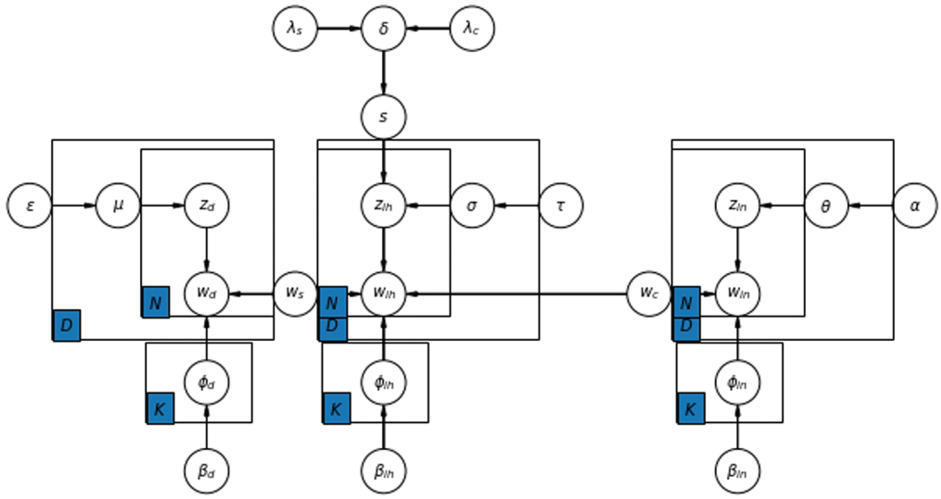

Figure 1. Plate diagram of ccTM.

Figure 1 shows the plate diagram of our use case of the ccTM model. We rephrase the ccTM generative process in our use case as described:

1. For each topic $k \in 1, \ldots, K$,
   a. Draw $\phi_{d,k}$ from the cited-specific topic-word distribution $\phi_{d,k} \sim Dir(\beta_d)$.
2. For each document $d \in 1, \ldots, D$,
   a. Draw θ_d from the document-topic distribution $\theta_d \sim Dir(\alpha)$ of the citation document.
   b. Draw μ_d from the document-topic distribution $\mu_d \sim Dir(\epsilon)$ of the citation document.
   c. Draw σ_d from the document-topic distribution $\sigma_d \sim Dir(\tau)$ of the citation document.
   For each topic, draw a Bernoulli distribution δ_k from Beta distribution
3. $\delta_k \sim Beta(\lambda_s, \lambda_c)$
4. For $w_{d,i}$, which is the ith word in document d:
   a. The random variable s_1, s_2 obey Bernoulli distribution $s_{d,i} \sim Bernoulli(\delta_k)$
   b. Draw $z_{d,d,i}$ from the multinomial distribution $z_{d,d,i} \sim Mult(\mu_d)$
   c. Draw $z_{ih,d,i}$ from the multinomial distribution $z_{ih,d,i} \sim Mult(\sigma_d)$
   d. Draw $z_{in,d,i}$ from the multinomial distribution $z_{in,d,i} \sim Mult(\theta_d)$
      i. If $s_1 = 1$:
         1. Draw $w_{s,d,i}$ from the multinomial distribution $w_{s,d,i} \sim Mult(\phi_{d,k})$.
      ii. If $s_1 = 0$:
         1. Draw $w_{s,d,i}$ from the multinomial distribution $w_{s,d,i} \sim Mult(\phi_{ih,k})$.
      iii. If $s_2 = 1$:
         1. Draw $w_{c,d,i}$ from the multinomial distribution $w_{c,d,i} \sim Mult(\phi_{in,k})$.
      iv. If $s_2 = 0$:
         1. Draw $w_{c,d,i}$ from the multinomial distribution $w_{c,d,i} \sim Mult(\phi_{ih,k})$.

We implement ccTM via the Python package PyMC (Salvatier, Wiecki & Fonnesbeck, 2016), which expresses this generation process in an intuitive way.

4 Demystifying the research dynamics of highly cited researchers

4.1 "Citations of citations" data collection and description

We use Semantic Scholar Open Corpus (Ammar et al., 2018) to extract forward citations of 280 of Geoffrey Hinton's publications. Publication and citation generations are hierarchical and tree structured. Publications are the root nodes r. First generation citations c_1 are child nodes of the publications. Second generation citations c_2 are child nodes of the first generation citations c_1. The dataset can be denoted as D = $\{r, c_1, c_2, \cdot, c_n\}$.

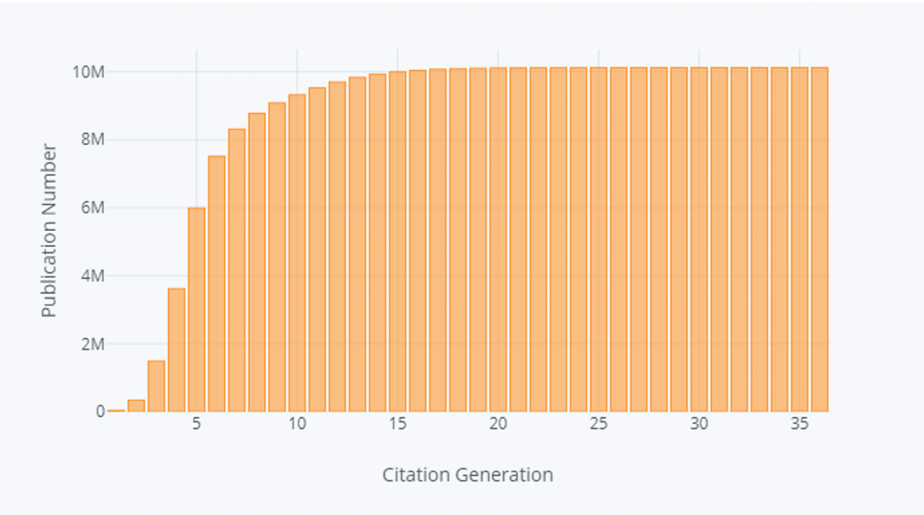

Figure 2. Publication number of each citation generation.

Since Semantic Scholar Open Corpus is a closed collection, we are able to finally extract all forward citation chaining. To reach a steady state for citation collection, we have two kinds of possible state: the next extraction loop is either terminated with no more citations, as shown in Equation (1) below. or the next extraction loop is the current data, which means it is a self-loop in terms of citation generations, as shown in Equation (2) below. We use *ss* to denote the steady-state. Then we determine the steady-state to stop citation extraction based on the following two criteria:

$$N_{c_{n+1}} = N_{c_{n+2}} = 0 \tag{1}$$

$$\frac{N_{c_{n+1}}}{N_{c_n}} = \frac{N_{c_{n+2}}}{N_{c_{n+1}}} \tag{2}$$

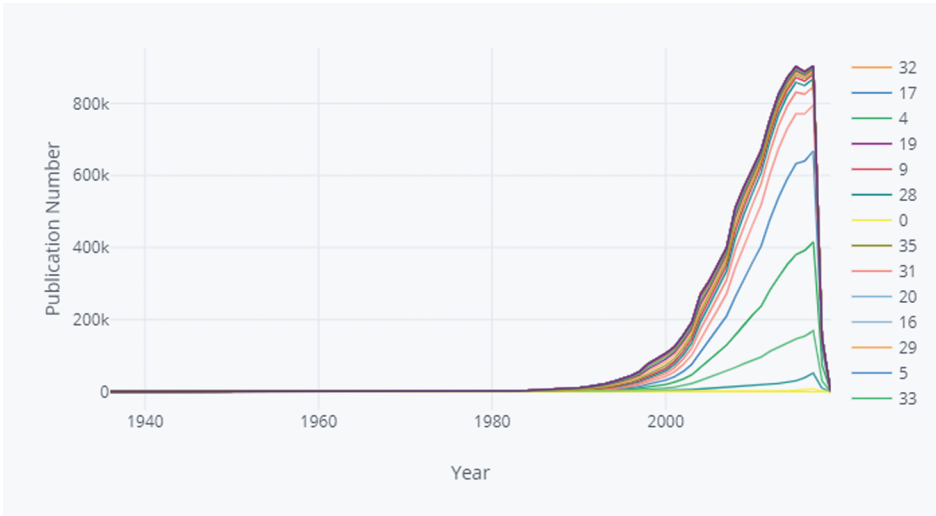

Figure 3. Yearly distribution for each citation generation and original cited paper.

Through exhaustively extracting citations of citations from Semantic Scholar Open Corpus, we obtain the steady state for Equation (2). This state is one in which all citations have become a closed network. We get 36 citation generations and 39,219,709 candidate documents in total. The publication number of each citation generation is depicted in Figure 2. We plot the yearly distribution of each generation of citations in Figure 3. In the legend, *s* refers to the original papers of Hinton, and the numbers represent the forward citation generation.

4.2 Data preprocessing

Though most topic models use bag-of-words representations, it is generally recognized that single words lack interpretability. Recent work by Cardenas et al. (2018) also reports that an entity can improve topic coherence compared to bag-of-words. In our work, we use entities that are provided by Semantic Scholar Open Corpus.

Part of the extracted citation dataset, however, is biology-related or electronics related. To limit the data from the artificial intelligence field, we limit our search to publications that contain any of the top 100 entities of Hinton's publications. After this refinement step, we get 341,570 publications and 1,520,389 cited connections, including 160 of Geoffrey Hinton's publications.

Since we want to compare the cited document and each of its citation generations, another important step is to map Hinton's original publication with every publication

of the citation generations. One citation instance may be in-different citation generations. To avoid repeatedly calculating one single citation, we only count newly add citations in each citation generation.

Since there are isolated cited-citing tuples after the previous refinement step, we only get 30 citation generations and 328,105 publications that have been attached to the original paper by Hinton. These cited-citing tuples are then aggregated with entities merged. Cited-inclusive words, shared words and citing-inclusive words are calculated for each cited publication. We describe Hinton's publication statistics over citation generation in Figure 4, and and we describe publication statistics of each citation generations in Figure 5.

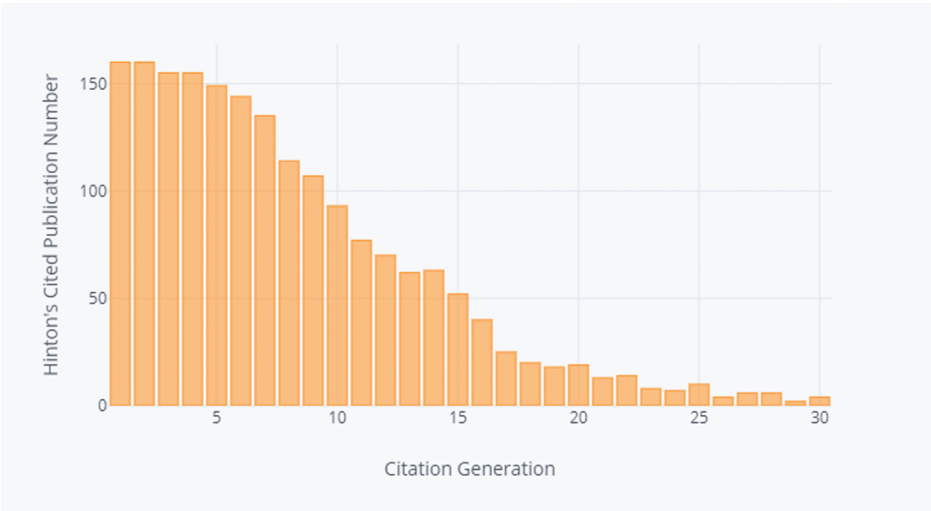

Figure 4. Hinton's cited publication number of each citation generation after preprocessing.

4.3 Topic dynamics

We apply the cross-collection topic model to the set of 30 cited-citing tuples. For each citation generation, we get topic disappearance, topic inheritance and topic innovation distribution. We illustrate the topic distribution results of the 8th citation generation in Figure 6, as well as the top 10 entities of each topic in Table 1.

The results show that the "Boltzmann" topic is the major disappearing topic. The "Markov" topic and the topic about "artificial neural networks" are the top-most inherited topics. The "numerical" topic is the most innovative topic. Due to limited space, we will not illustrate all results of each citation generation.

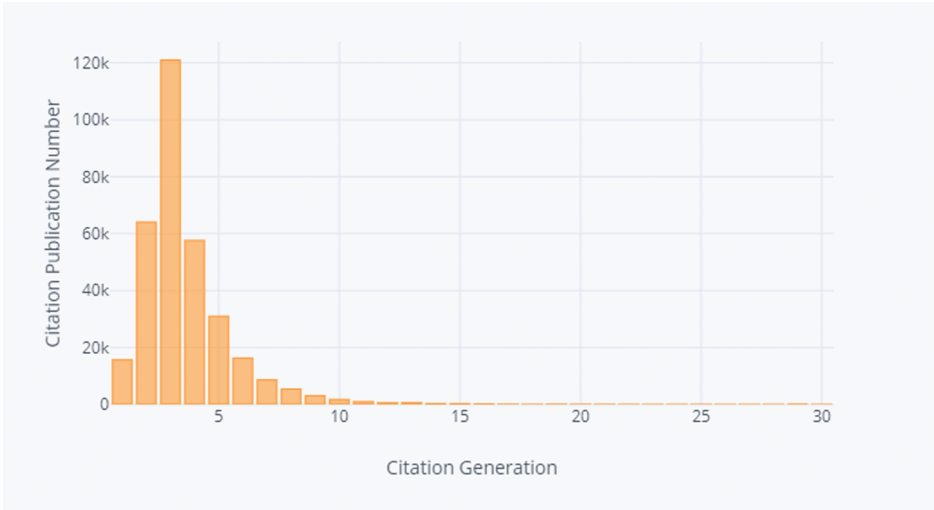

Figure 5. Publication number of each citation generation after preprocessing.

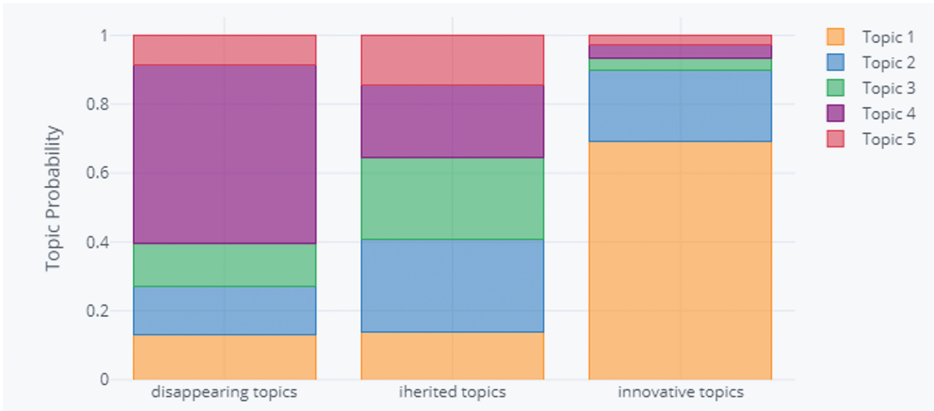

Figure 6. Disappearing topic, inherited topic, and innovative topic distributions of the 8th citation generation.

The main aim of this part is to see the dynamics of disappearing topics, inherited topics and innovative topics over citation generations. In the following section, we list the top topic entities among disappearing, inherited and innovative topics. Though we have 30 citation generations, the generations after generation 22 only contain less than 30 publications, which is not sufficient to draw meaningful conclusions. Thus we only include 22 citation generation topic results in this paper.

Table 1. Topic-entity distribution of the 8th citation generation.

Topic	Top 10 Topic Entities
1	numerical analysis, simulation, Monte Carlo, artificial intelligence, dynamic programming, probability, principal component analysis, experiment, Markov chain, controllers
2	algorithm, simulation, Markov chain, Monte Carlo method, Monte Carlo, artificial intelligence, principal component analysis, experiment, program optimization, artificial neural network
3	artificial neural network, algorithm, fingerprint, genetic programming, biological neural networks, CPU cache, backpropagation, neural network simulation, gradient, discontinuous Galerkin method
4	artificial neural network, Boltzmann machine, restricted Boltzmann machine, generative model, backpropagation, pixel, speech recognition, deep learning, MNIST database, mixture model
5	fault tolerance, data mining, artificial neural network, brute force search, algorithm, asymptotically optimal algorithm, backpropagation

Table 2. Disappearing topics over each citation generation.

Gen	Top 10 Disappearing Topic Entities
1-2	generative model, Boltzmann machine, restricted Boltzmann machine, algorithm, inference, pixel, latent variable, gradient, Markov chain, approximation algorithm
3-18	artificial neural network, algorithm, generative model, backpropagation, nonlinear system, deep learning, gradient, speech recognition, hidden Markov model, pixel
19	artificial neural network, generative model, machine learning, algorithm, restricted Boltzmann machine, convolutional neural network, image resolution, value ethics, Boltzmann machine, gradient
20	artificial neural network, hidden Markov model, Markov model, nonlinear system, backpropagation, unsupervised learning, speech recognition, time series, cluster analysis, cognition disorders
21	artificial neural network, nonlinear system, generative model, factor analysis, MNIST database, anatomical layer, deep learning, mixture model, unit, gradient
22	pixel, restricted Boltzmann machine, gradient, artificial neural network, speech recognition, Boltzmann machine, unsupervised learning, statistical model, deep learning, network architecture

4.3.1 Disappearing topic dynamics

We list the top 10 disappearing topic entities in Table 2, in which several citation generations are merged with regards to topics with similar meanings. For direct citations and first indirect citations, the same disappearing topics are related to "generative" and "Boltzmann." Citation generations 3 through 18, 20, and 21 abandon Hinton's notions of "backpropagation" and "deep learning." Citation generations 19 and 22 also abandon topics related to "generative" and "Boltzmann."

The indirect citations of citation generations 2 to 22 should contain all of Hinton's notions, because they do not directly cite Hinton's publications. Topics like "generative" and "Boltzmann" are kept by several middle citation generations. Topics like "backpropagation" and "deep learning" are kept by several end citation generations. This, however, is more explicit in the inherited topics results.

Table 3. Inherited topics over each citation generation.

Gen	Top 10 Inherited Topic Entities
1-7	artificial neural network, algorithm, deep learning, backpropagation, speech recognition, hidden Markov model, neural network simulation, machine learning, test set, nonlinear system
8-9	algorithm, simulation, Markov chain, Monte Carlo method, Monte Carlo, artificial intelligence, principal component analysis, experiment, program optimization, artificial neural network
10-14	artificial neural network, backpropagation, generative model, Boltzmann machine, restricted Boltzmann machine, computer data storage, deep learning, speech recognition, feedforward neural network, nonlinear system
15-16	simulation, Monte Carlo method, Monte Carlo, algorithm, numerical analysis, Markov chain, dynamic programming, solutions, coefficient, experiment
17	artificial neural network, gradient, matching polynomial, nonlinear system, spline interpolation, hidden Markov model, generative model, approximation algorithm, Bayesian network, factor analysis
18	simulation, Monte Carlo method, Monte Carlo, computation, computation action, silicon, gradient, distortion, Markov chain, algorithm
19	artificial neural network, generative model, machine learning, algorithm, restricted Boltzmann machine, convolutional neural network, image resolution, value ethics, Boltzmann machine, gradient
20	artificial neural network, hidden Markov model, Markov model, nonlinear system, backpropagation, unsupervised learning, speech recognition, time series, cluster analysis, cognition disorders
21	artificial neural network, nonlinear system, generative model, factor analysis, MNIST database, anatomical layer, deep learning, mixture model, unit, gradient
22	artificial intelligence, mitral valve prolapse syndrome, greater than, power dividers and directional couplers, supervised learning, performance, meal occasion for eating, plasminogen activator, nominal impedance, platelet glycoprotein 4 human

4.3.2 Inherited topic dynamics

The dynamics of inherited topics are shown in Table 3. Citation generations 1 to 7 and 10 to 14 all inherit topics related to "artificial neural network" and "backpropagation." Citation generations 2 to 7 and 10 to 14, however, do not directly cite Hinton's work, even though "backpropagation" was first proposed by Hinton. Citation generations 8 to 9, 15 to 16, and 18 all inherit topics related to "Monte Carlo." Citation generation 19 inherits topics related to "Boltzmann" without directly citing Hinton's publications. The remaining citation generations inherit topics related to "artificial neural network" and "Markov," which were not invented by Hinton. The topics of "artificial neural network" and "backpropagation" are certainly the most lasting impact of Hinton's academic contribution. Even without being directly cited, these topics saw an impact that lasted until the 14th citation generation. Another unpopular topic, "Boltzmann," also inspired work during 19th citation generation. "Monte Carlo," however, was not a part of Hinton's original work, so its inherited topics here are actually shared topics.

The disappearing topics and inherited topics, however, are not complementary to each other in this model. This is because we model the document-wise observation of topics in 160 of Hinton's publications. Some disappearing topics might also be inherited topics of the other documents, as well.

Table 4. Innovative topics over each citation generation.

Gen	Top 10 Innovative Topic Entities
1	artificial intelligence, computation, machine learning, biological neural networks, experiment, neural tube defects, convolutional neural network, synthetic data, simulation, neural networks
2	machine learning, experiment, supervised learning, simulation, program optimization, sparse matrix, neural networks, neural network simulation, computation, unsupervised learning
3	greater than, solutions, classification, estimation theory, Eisenstein's criterion, pattern recognition, cluster analysis, neural tube defects, feature selection, sensor
4	robot, Monte Carlo, Markov model, Eisenstein's criterion, rule guideline, neural network simulation, coefficient, numerical analysis, dynamic programming, high and low level
5	numerical analysis, artificial intelligence, heuristic, experiment, solutions, Eisenstein's criterion, computation, requirement, sensor, coefficient
6-21	artificial intelligence, Monte Carlo method, biological neural networks, neural network simulation, Bayesian network, Markov chain
22	principal component analysis, food, principal component, obesity, platelet glycoprotein 4 human, red meat, whole grains, eaf2 gene, diabetes mellitus, exercise

4.3.3 Innovative topic dynamics

The innovative topic dynamics of each citation generation are shown in Table 4. Each innovative topic of different citation generations varies in entities. There is no obvious pattern to the dynamics. Citation generations 1 to 2, however, are innovative in topics related to "artificial intelligence" and "machine learning." "Monte Carlo" is the main innovative topic of citation generations 4 and 6 to 21. Citation generations 3 and 5 are both innovative in topics related to "numeric." On the other hand, citation generation 22 contains topics related to "health" and "physical."

5 Conclusion and future work

Our work conducts citation content analysis, and it can provide powerful insights into research on dynamic evolution. Though not explicitly studied, this model can also be used for plagiarism detection if a large portion of inherited topics exist without direct citation. In this paper, we map topic dynamics through each citation level. Some disappearing topics appear on and off through the dynamic routines of the research, and some inherited topics lasted long, with new innovative topics emerging as well.

Finally, though our model reveals insights into the research dynamics of citation generations, our implementation of the ccTM model is still slow for each citation iteration. In future work, we will try batched variational inference, which is said to be faster than Gibbs sampling. The other direction to explore is how to obtain more robust topics through topic modeling.

Acknowledgement

This work is supported by the Programs for the Young Talents of National Science Library, Chinese Academy of Sciences (Grant No. 2019QNGR003).

Author contributions

Xiaoli Chen (chenxl@mail.las.ac.cn): Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Wrote the paper. Tao Han (hant@mail.las.ac.cn): Conceived and designed the analysis.

References

Ammar, W., Groeneveld, D., Bhagavatula, C.S., Beltagy, I., Crawford, M., Downey, D.C., & Dunkelberger, J. (2018). Construction of the literature graph in semantic scholar. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 3, pp. 84–91. United States: Association for Computational Linguistics (ACL). doi:10.18653/v1/N18-3011

Bao, Y., Collier, N., & Datta, A. (2013). A partially supervised cross-collection topic model for cross-domain text classification. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 239–248. New York, USA: ACM. doi:10.1145/2505515.2505556

Beykikhoshk, A., Phung, D., Arandjelovic, O., & Venkatesh, S. (2016). Analysing the history of autism spectrum disorder using topic models. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 762–771. Montreal: IEEE. doi:10.1109/dsaa.2016.65

Blei, D.M., Ng, A.Y., Jordan, M.I., & Lafferty, J. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Cardenas, R., Bello, K., Coronado, A.M., & Villota, E. (2018). Improving topic coherence using entity extraction denoising. Prague Bull. Math. Linguistics, 110, 85–101. doi: 10.2478/pralin-2018-0004

Chang, J. (2009). Relational topic models for document networks. In Proceedings of the Conference on AI and Statistics (AISTATS).

Chang, J., & Blei, D.M. (2010). Hierarchical relational models for document networks. Annals of Applied Statistics, 4(1), 124–150.

Chen, C., Buntine, W., Ding, N., Xie, L., & Du, L. (2015). Differential topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 230–242. doi:10.1109/TPAMI.2014.2313127

Chen, X., & Han, T. (2019). How research milestone shape the technology of today—A case study of highly cited researcher using topic model. In Proceedings of the 17th International Conference on Scientometrics and Informetrics, ISSI 2019, pp. 2553–2554. Rome.

De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. Scientometrics, 103, 413–433. doi:10.1007/s11192-015-1554-1

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. ICML '07: In Proceedings of the 24th International Conference on Machine Learning, pp. 233–240. Retrieved from https://doi.org/10.1145/1273496.1273526

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, pp. 281–288. New York, USA: ACM. doi:10.1145/1553374.1553410

Elgendi, M. (2019). Characteristics of a highly cited article: A machine learning perspective. IEEE Access, 7, 87977–87986. doi:10.1109/ACCESS.2019.2925965

Gerrish, S.M., & Blei, D.M. (2010). A language-based approach to measuring scholarly impact. In Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 375–382. USA: Omnipress. Retrieved from http://dl.acm.org/citation.cfm?id=3104322.3104371

Hall, D., Jurafsky, D., & Manning, C.D. (2008). Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 363–371. Stroudsburg: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1613715.1613763

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C. (2009). Detecting topic evolution in scientific literature: How can citations help? pp. 957–966. doi:10.1145/1645953.1646076

Hu, X., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. Journal of Informetrics, 5, 27–36. doi:https://doi.org/10.1016/j.joi.2010.07.004

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online multiscale dynamic topic models. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 663–672. New York, USA: ACM. doi:10.1145/1835804.1835889

Jennifer, S., & Halem, M. (2018). Ontology-grounded topic modeling for climate science research. In Emerging Topics in Semantic Technologies. ISWC 2018 Satellite Events. AKA Verlag, Berlin.

Kataria, S., Mitra, P., & Bhatia, S. (2010). Utilizing context in generative bayesian models for linked corpus. In M. Fox, & D. Poole (Ed.), In Proceedings of the 24th AAAI Conference on Artificial Intelligence. AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1883

Kim, J., Kim, D., & Oh, A. (2017). Joint modeling of topics, citations, and topical authority in academic corpora. Transactions of the Association for Computational Linguistics, 5, 191–204. Retrieved from https://transacl.org/ojs/index.php/tacl/article/view/1061

Li, W., & Mccallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd International Conference on Machine Learning, pp. 577–584.

Martínez, M.A., Herrera, M., Contreras, E., Ruíz, A., & Herrera-Viedma, E. (2015). Characterizing highly cited papers in Social Work through H-Classics. Scientometrics, 102, 1713–1729. doi:10.1007/s11192-014-1460-y

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2, pp. 3111–3119. USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999792.2999959

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272. Stroudsburg: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145462

Moody, C.E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec.

Musat, C.C., Velcin, J., Trausan-Matu, S., & Rizoiu, M.-A. (2011). Improving topic evaluation using conceptual knowledge. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain.

Nallapati, R.M., Ahmed, A., Xing, E.P., & Cohen, W.W. (2008). Joint latent topic models for text and citations. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 542–550. New York, USA: ACM. doi:10.1145/1401890.1401957

Nallapati, R., & Cohen, W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence in blogs. International Conference on Weblogs and Social Media.

Newman, D., Chemudugunta, C., & Smyth, P. (2006). Statistical entity-topic models. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 680–686. New York, USA: ACM. doi:10.1145/1150402.1150487

Newman, D., Lau, J.H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Stroudsburg: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1857999.1858011

Parker, J.N., Allesina, S., & Lortie, C.J. (2013). Characterizing a scientific elite (B): Publication and citation patterns of the most highly cited scientists in environmental science and ecology. Scientometrics, 94(2), 469–480. doi:10.1007/s11192-012-0859-6

Paul, M., & Girju, C.R. (2009). Topic modeling of research fields: An interdisciplinary perspective. International Conference Recent Advances in Natural Language Processing, RANLP, 337–342.

Paul, M., & Girju, R. (2010). A two-dimensional Topic-Aspect Model for discovering multi-faceted topics. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 545–550.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. Valletta: ELRA.

Risch, J., & Krestel, R. (2018, 6). My approach = Your apparatus? Entropy-based topic modeling on multiple domain-specific text collections. In Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries. Fort Worth, TX, USA. doi:10.1145/3197026.3197038

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408. New York, USA: ACM. doi:10.1145/2684822.2685324

Salvatier, J., Wiecki, T., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. PeerJ Computer Science, e55. doi:10.7287/PEERJ.PREPRINTS.1686V1

Shen, J., Song, Z., Li, S., Tan, Z., Mao, Y., Fu, L., . . . , & Wang, X. (2016). Modeling topic-level academic influence in scientific literatures. Scholarly big data: AI perspectives, challenges, and ideas, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12598

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. Tech. rep. Retrieved from https://www.microsoft.com/en-us/research/publication/continuous-time-dynamic-topic-models/

Wang, X., Zhai, C., & Roth, D. (2013). Understanding evolution of research themes: A probabilistic generative model for citations. In R. Parekh, J. He, D. S. Inderjit, P. Bradley, Y. Koren, R. Ghani, . . . R. Uthurusamy (Ed.), KDD 2013 - 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1115–1123. Association for Computing Machinery. doi:10.1145/2487575.2487698

Wu, H., Wang, M., Feng, J., & Pei, Y. (2010). Research topic evolution in "Bioinformatics". In Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering, pp. 1–4. doi:10.1109/ICBBE.2010.5516318

Wu, Q., Zhang, C., Hong, Q., & Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. Journal of Information Science, 40(5), 611–620. doi:10.1177/0165551514540565

Xu, S., Shi, Q., Qiao, X., Zhu, L., Jung, H., Lee, S., & Choi, S.P. (2014). Author-Topic over Time (AToT): A dynamic users' interest model. In J. J. Park, H. Adeli, N. Park, & I. Woungang (Ed.), Mobile, Ubiquitous, and Intelligent Computing, pp. 239–245. Berlin: Springer Berlin Heidelberg.

Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis: Research Dynamics, Impact, and Dissemination. Journal of the Association for Information Science and Technology, 66, 2357–2372. doi:10.1002/asi.23324

Zhai, C., Velivelli, A., & Yu, B. (2004). A cross-collection mixture model for comparative text mining. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 743–748. New York, USA: ACM. doi:10.1145/1014052.1014150

Zhang, J., Gerow, A., Altosaar, J., Evans, J., & Jean So, R. (2015). Fast, flexible models for discovering topic correlation across weakly-related collections. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1554–1564. Lisbon: Association for Computational Linguistics. doi:10.18653/v1/D15-1179

Zhou, H.K., Yu, H.M., & Hu, R. (2017). Topic discovery and evolution in scientific literature based on content and citations. Frontiers of Information Technology & Electronic Engineering, 10, 1511–1532. doi:10.1631/FITEE.1601125