Collecting the Seminal Scientific Abstracts with Topic Modelling, Snowball Sampling and Citation Analysis

Hennadii Dobrovolskyi, Nataliya Keberle

Department of Computer Science, Zaporizhzhya National University,
Zhukovskogo st. 66, 69600, Zaporizhzhya, Ukraine,
gen.dobr@gmail.com, nkeberle@gmail.com

Abstract. This paper presents a complete information technology for collecting and analysis of a citation network of scientific publications aimed at detecting of seminal papers in a selected domain of research. The technology consists of the seed paper selection, plain snowball sampling, probabilistic topic modeling, greedy restricted snowball sampling, and analysis of the collected citation network.
The topic model is built on the base of word-word co-occurrence probability with combination of sparse symmetric nonnegative matrix factorization and principal component approximation. Experiments with the collection of High Energy Physics abstracts show that the number of topics in the model is determined in natural way and the Kullback-Leibler divergence correlates with cosine similarity calculated from keywords provided by publication authors.
The citation networks on "critical thinking" and "automatic pronunciation assessment" domains are collected and analyzed. The analysis shows that both networks are "small worlds" and therefore the observed saturation of the restricted snowball sampling can provide the complete set of publications in domains of interest. Multiple runs of the sampling confirm the hypothesis that the set of seminal publications is stable with respect to variations of the seed papers. The modified main path analysis allows to distinguish the seminal papers including new publications following main stream of research.

Keywords: text mining, short text document, topic modelling, principal component analysis, sparse symmetric nonnegative matrix factorization, citation network, main path analysis.

1 Introduction

The quality of a related work review is a problem well known to each scientist. The questions to be answered are "If the collected scientific publications contain all notable scientific results of the domain of interest?" and "Which of the collected publications direct the mainstream of the knowledge evolution in the domain of interest?" Below we present a complete information technology for

getting answers to both questions. The essence of the technology is the collecting and analysis of a citation network of scientific publications aimed at detecting of seminal papers in a selected domain of research.

To our best knowledge, while the separate parts of the method are developed, the entire procedure that takes the small set of papers on some scientific domain and produces perfect list of references is not known.

The objectives of the presented work are:

- to present the complete technology that takes a manually selected seed papers and produces the short list of interconnected scientific publications that reflects evolution of main ideas of the selected scientific domain. The technology contains both restricted snowball sampling method and citation network analysis.
- to test all initial assumptions, namely
  - if the proposed snowball restriction method [6] provides adequate semantic distance between publications;
  - if the obtained restricted snowball forms the scale-free network [22];
  - if the restricted snowball provides saturation of the publication dataset;
  - if the best age of the seed papers is 5–10 years;
  - if the biased seed papers can produce unbiased citation network.

The distinctive features of the presented method are application of the probabilistic topic model to perform restricted snowball sampling and collect citation network, then the main path analysis is applied to point both the most influential publications and the main path of scientific knowledge evolution. The main path allows detecting the newest publications that follow the mainstream and the outliers that potentially contain the completely new ideas.

The structure of the paper is following. Section 2 overviews the publications related to the presented technology, Section 3 contains description of the crucial steps of the algorithm, Section 4 states the experiment pre-conditions and Section 5 discusses the results. Conclusion summarises the main results and discusses future work.

2 Related Work

Publications on a domain of knowledge can be collected from conference proceedings [7], the study of the co-authorship [15, 16], elaboration of keywords and topics [14, 17], querying academic search engines[1] with a set of keywords or snowball sampling [10, 1, 6]. However, not for every research domain there is a corresponding conference, as well as one author can write papers on different topics. Building maps and ontologies of large scientific domains does not provide the list of references rather the set of interconnected concepts. Querying with

[1] Google Scholar, https://scholar.google.com,
Semantic Scholar, https://www.semanticscholar.org/,
Microsoft Academic, https://academic.microsoft.com/

a set of keywords produces the biased set of publications [19] because different researchers use slightly different terms to report their results.

The most appropriate way to collect scientific papers is snowball sampling [10, 1], when each publication from the current queue is considered then all referenced papers and all papers referencing to the publication are added to the next level queue. The snowball sampling allows collecting publications on the narrow research topic and connect them in the citation network [22]. The high quality of citation-based search algorithm is provided with phenomena of "small world" which is a proved property of scale-free networks [2, 22]. Newman [15] has shown that in the most of the cases it is enough to do three iterations. However, the statistical properties of global citation network including all scientific papers is not known, that is why we need to test if the small world assumption is true for the collected subset of citation network and if the three iterations allows collecting most of the papers.

Another point of snowball sampling is dependence on the initial queue called a seed collection. The general advice [10] recommends that the seed papers should be the seminal papers of the knowledge domain pointed by experts or the papers selected by the researcher. Valid seed papers should be 5–10 years old and have to be widely cited. The best seeds are the reviews, foundational or framing articles on the topic of interest. However, the advice also should be checked. Moreover, we need to test if the biased seed papers can produce unbiased citation network.

The snowball sampling cannot be applied directly to publication crawling because the list of references can contain the items that are not directly related to the investigated domain. Therefore the straightforward implementation of snowball publication sampling causes infinite collection inflation and some restrictions should be introduced to accept or reject the candidate publication. It should be noted that the introduced restrictions can violate the small world property and we need to check if the restricted snowball result is scale-free citation network.

To filter out the most relevant papers while sampling Ahad et al. [1] in their approach use vector document model and cosine similarity, however the document vector model relies on word spelling rather than meaning that causes precision loss when the short texts are considered. Lecy et al. [10] apply PageRank calculated by Google Scholar as a measure of paper significance. However, PageRank is a property of a global citation network including all topics of knowledge, so it cannot be calculated from its small subset. One of the most promising approaches is the probabilistic topic model (PTM).

Probabilistic topic models [24] use a large collection of documents and statistical approach to model words and documents as vectors in a high-dimensional semantic space R^n, where n is much less than number of words and number of documents. The base idea of PTM is to construct few topis which are groups of tightly connected words. Then document words are represented as a result of two-stage random sampling. The most known method of topic modelling is Latent Dirichlet Allocations (LDA) [5] which is successful and simple enough.

A general introduction and survey of the topic modelling can be found in [24] along with a novel approach, called Additive Regularization of Topic Models.

However, in most of the scientific databases, full texts are often protected by copyright. Therefore the only information we can use are paper title, paper abstract, and sometimes the database-specific keywords and topics. So the documents that we analyse are short and common PTMs based on document-word statistics lose their precision. This shortcoming is overpassed with approaches utilizing word co-occurrence statistics in Biterm Topic Model (BTM) [25] and Word Network Topic Model (WNTM) [26] instead of counting document-word pairs. Another method of word embedding, called GloVe, is proposed in [18]. It is based on word-word co-occurrence matrix and uses global matrix factorization, so it is close to BTM [25] and WNTM [26] statistical topic modelling.

Also, the vague part of common PTMs is that number of topics cannot be determined with document analysis. To overcome this weakness, handling the word-word co-occurences with principal component analysis (PCA) and sparse symmetric nonnegative matrix factorization (Sparse SNMF) was proposed [6].

The collected citation network [22] can be analyzed using citation count and other simple statistics [12], PageRank [11], information retrieval techniques [1], knowledge graph [17], combined supervised machine learning approaches [23] or Main Path analysis [12].

The most appropriate way to highlight the seminal papers of the small scientific domain is main path analysis because the method deals only with the collected citation network and allows to increase the precision of sampled dataset. On the contrary, the citation count and other statistics, supervised machine learning applied by Valenzuela, Ha and Etzioni [23] and PageRank cannot point out the tightly interconnected subset of the citation network. Klink-2 [17] and similar algorithms aim to build the map of knowledge domain but do not seek the most influenced publications.

3 Information Technology Overview

The general workflow of the restricted snowball sampling is introduced in [6]. It contains the following steps:

1. Collect a set of seed papers and put them in the initial, 0-th, queue.
2. Run several iterations of the unrestricted snowball sampling to pickup baseline documents. For $n \in 0, 1, 2, 3$
   1 get a portion of papers from the n-th queue;
   2 download the papers referenced by the portion;
   3 download the papers referencing the portion;
   4 add all the downloaded papers to the $(n + 1)$-th queue.
3. Create the PTM using baseline documents:
   1 extract title and abstract from each document of the collection;
   2 split all the titles and abstracts into sentences;
   3 create the dictionary containing all the nouns and adjectives that occur in the sentences;

    4 combine all terms from the reduced dictionary occurring in the same sentence into pairs and build the joint probability matrix;
    5 detect the collection specific stop-words and exclude them from the reduced dictionary;
    6 perform Sparse SNMF to create PTM;
    7 map each of the seed papers to a vector of topic probabilities.
4. Perform the batch restricted snowball sampling: for $n \in 0, 1, 2, 3$
    1 get a portion of papers from the n-th queue;
    2 download the papers referenced by the portion;
    3 download the papers referencing the portion;
    4 extract bag of stemmed words from each of downloaded papers;
    5 map each of the downloaded papers to a vector of topic probabilities;
    6 calculate distance from each downloaded paper to the seed papers;
    7 add to the next level queue only those of downloaded papers which are close to the seed papers.
5. Analyse the citation network.

The details of the restricted snowball sampling and probabilistic topic model construction are discussed in [6].

4 Citation Network Analysis

4.1 Cycles Elimination

The correctly built citation network must be an acyclic directed graph. However, the publication database errors accidentally can cause cycles. The problem with cycles is that if there is a cycle in a network then there is also an infinite number of paths between some vertices. Since a citation network is usually almost acyclic to transform it into an acyclic network we use the "preprint" transformation described by Batagelj [3]. First, we identify cycles and then each paper from a cycle is duplicated with its "preprint" version and the papers inside cycle cite "preprints".

4.2 Simple Citation Path Count

Our approach is similar to Search Path Count (SPC) algorithm [12, 3]. We introduce two pseudo-vertices – source and target. A vertex that does not reference any other publication vertex, gets an edge to the target vertex. A vertex that is not referenced by any publication vertex, gets an edge from the source vertex, so the graph becomes connected. Next step is to calculate all simple paths from the source to the target using Python library NetworkX[2]. The algorithm uses a modified depth-first search to generate the paths [9]. As the result we obtain a set of paths, each of which is a sequence of vertices. Each pair of direct neighbour vertices in such a sequence is an edge in the citation graph. For each edge, its

[2] NetworkX, https://networkx.github.io

frequency is calculated against all paths – a number of paths through it, simple path count. Next we calculate edge resistance as inverse proportional to edge simple path count – this allows diminishing the difference among the most cited and least cited papers [21, 8]. The path resistance is then calculated as the sum of its edge resistances. Finally, we set an order over the paths using the path resistances. Using path resistances is a distinguishing feature of the proposed algorithm.

The difference of the applied algorithm from SPC algorithm is the preservation of the citation graph connectivity. In the known algorithm [4], as soon as the edge SPC scores are calculated the edges having low scores are removed and the citation network can become a disconnected graph.

4.3 Chasing New Ideas in Publications

Path resistance allows detecting new publications in the field, not referenced yet by any other authors but existing in a mainstream of the domain. We can separate all papers into mainstream research and probably new research fields or directions. The smallest (up to a certain threshold) path resistances correspond to the mainstream, whereas the biggest path resistances correspond to the publications that are either brand new, bad written or published in a low impact journal/conference proceedings. We assume those publications are the source of potentially new ideas and topics.

5 Analysis of Experimental Results

5.1 Experimental Settings

We took three different corpora: "high energy physics"(HEP), "critical thinking"(CT) and "pronunciation quality assessment"(PQA).

HEP publications [20] are available from the European Laboratory for Nuclear Research. The hep-ex partition of the HEP collection is composed of 2802 abstracts related to experimental high-energy physics that are indexed with 1093 main keywords (the categories), the hep-astroph partition contains 2716 abstracts from astrophysics section and 18114 abstracts on theoretical physics in hep-th metadata. Each publication is manually annotated with keywords.

CT corpus is gathered with our snowball sampling software. The CT domain is characterized with a large noisy publications corpus tightly entangled with publications on psychology, didactics, pedagogy and phylosophy. The size of CT corpus is 24040 publication abstracts.

PQA domain is very specific and narrow, with a moderate-size corpus containing 8339 scientific abstracts collected by our snowball sampling software.

The sampling was run with following parameters: percentage of stop words to exclude – 2%; percentage of rare words to exclude – 5%; number of components in PCA which is maximal number of topics – 200; threshold KL-divergence – 0.18; sparsity parameter – 0.05; number of top citation paths – 50; minimal number of citations – 3.

5.2 Seminal Publications for PQA domain

Figure 1 shows the results of citation network analysis for PQA domain.

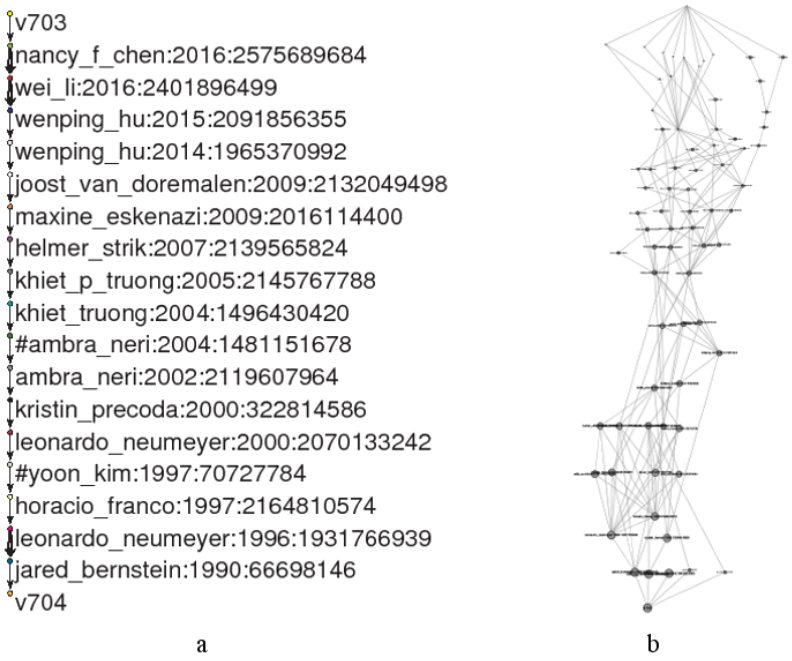

Fig. 1. Top path in PQA citation network (a) and top 73 paths of the citation network (b). Nodes are marked as (first author:year:MS_Academic_Id)

On the part (a) of Figure 1 we can see that the mainstream of pronunciation assessment contains the publications:

1. "Automatic evaluation and training in English pronunciation" by Bernstein, Cohen, Murveit, Rtischev, and Weintraub, 1990
2. "Automatic text-independent pronunciation scoring of foreign language student speech" by Neumeyer, Franco, Weintraub, and Price, 1996
3. "Automatic pronunciation scoring for language instruction" by Franco, Neumeyer, Kim, and Ronen, 1997
4. "Automatic pronunciation scoring of specific phone segments for language instruction" by Kim, Franco, and Neumeyer, 1997
5. "Automatic scoring of pronunciation quality" by Neumeyer, Franco, Digalakis, and Weintraub, 2000
6. "Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability" by Precoda, Halverson, and Franco, 2000

7. “The pedagogy-technology interface in computer assisted pronunciation training” by Neri, Cucchiarini, Strik, and Boves, 2002
8. “Segmental errors in Dutch as a second language: how to establish priorities for CAPT” by Neri, Cucchiarini, and Strik, 2004
9. “Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach” by Truong, 2004
10. “Automatic detection of frequent pronunciation errors made by L2-learners” by Truong, Neri, Wet, Cucchiarini, and Strik, 2005
11. “Comparing classifiers for pronunciation error detection” by Strik, Truong, Wet, and Cucchiarini, 2007
12. “An overview of spoken language technology for education” by Eskenazi, 2009
13. “Automatic detection of vowel pronunciation errors using multiple information sources” by Van Doremalen, Cucchiarini, and Strik, 2009
14. “A new neural network based logistic regression classifier for improving mispronunciation detection of L2 language learners” by Hu, Qian, and Soong, 2014
15. “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers” by Hu, Qian, Soong, and Wang, 2015
16. “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling” by Li, Siniscalchi, Chen, and Lee, 2016
17. “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning” by Chen and Li, 2016

The mainstream evolution of pronunciation assessment starts from application of automatic speech recognition, pays some attention to pedagogical aspects, goes to simple machine learning approaches, then to neural networks and to deep learning. Some of the seminal publications are the reviews containing discussions of the feature selection, methods comparison and combination. The part (b) of Figure 1 shows that the more top paths we keep, the more detailed knowledge map we obtain.

5.3 Assumptions Checking

PTM as a restriction criteria for the proposed restricted snowball sampling method provides adequate semantic distance between publications. To check the statement we took HEP collection, built PTM for it, and measure similarity using keywords annotating each publication from HEP collection. For each pair of HEP publications were calculated both symmetric KL-divergence and cosine similarity. The results are shown in Figure 2.

As we can see, the PTM-based symmetric Kullback-Leibler divergence [13] provides reliable upper bound for the keyword based cosine similarity. The reason is that the cosine similarity uses only word spelling and PTM uses R^n word embedding taking into account the meaning of terms.

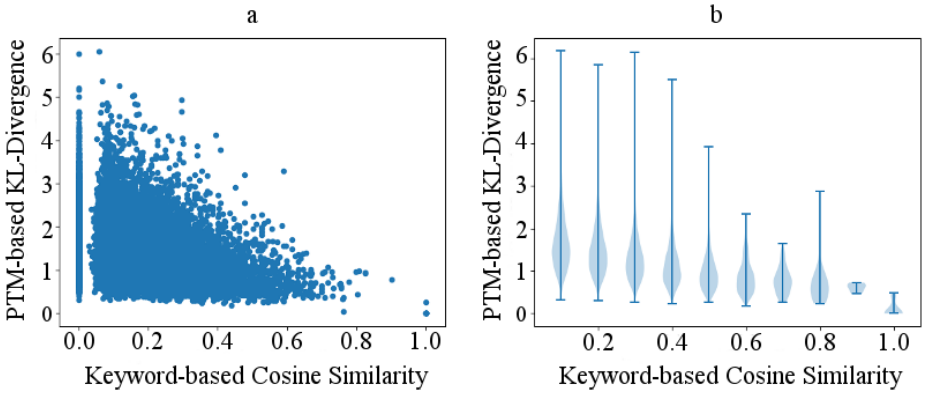

Fig. 2. Symmetric PTM-based KL-divergence and cosine similarity for the HEP abstracts: (a) scatter plot; (b) violin plot.

The restricted snowball sampling forms the scale-free network. Derek de Solla Price showed in 1965 [22] that the number of references to a paper (node degree) in a citation network had a heavy-tailed distribution following the power law and thus that the citation network is scale-free. One of our initial assumptions was that the restricted snowball sampling results in a scale-free networks so we need a few number of snowball iterations to achieve the high recall. Figure 3 was calculated on the base of PQA corpus and shows that for the small node degrees the logarithm of node number has linear dependency on the logarithm of node degree and for the large node degrees the dependency has heavy tail. That means, the restricted snowball sampling produces scale-free citation network as well as classical snowball. So we can be sure that a few iterations of the restricted snowball sampling allow collecting most of the relevant publications.

Saturation of the restricted snowball sampling. The restricted snowball sampling can be modelled as Poisson process when the publications appear sequentially and we can either (a) accept n-th publication and add it to the snowball or (b) don't accept. So we can calculate the confidence interval of Poisson distribution of event (a) and compare its upper bound with some pre-defined acceptance probability. Figure 4 shows 0.95 confidence interval of Poisson distribution of paper acceptance as a colored strip and acceptance probability threshold 0.05 as a straight line. The confidence interval was calculated on the base of 10 snowball runs for CT collection starting from random subsets of seed paper collection. After some number of tested abstracts the upper bound of confidence interval becomes lower than the threshold so the restricted snowball sampling guarantees the saturation.

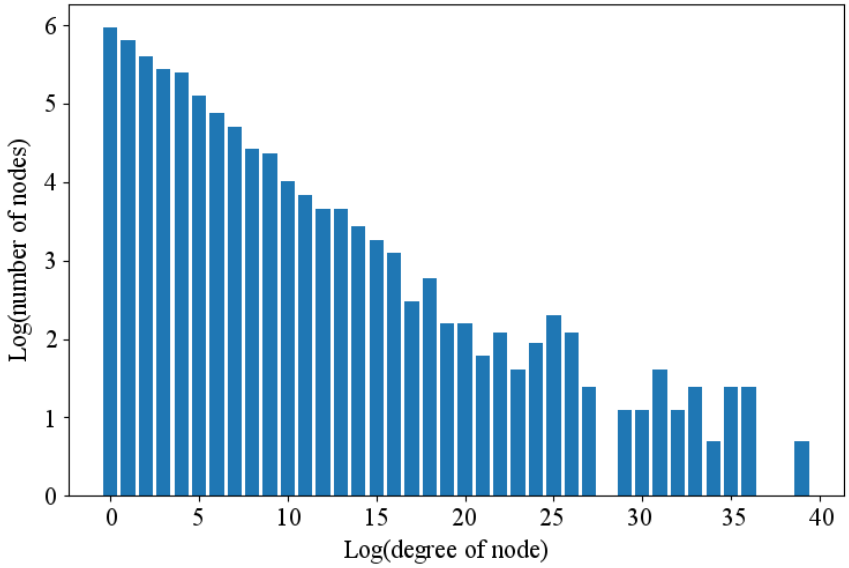

Fig. 3. Citation network node degree distribution in log-log scale.

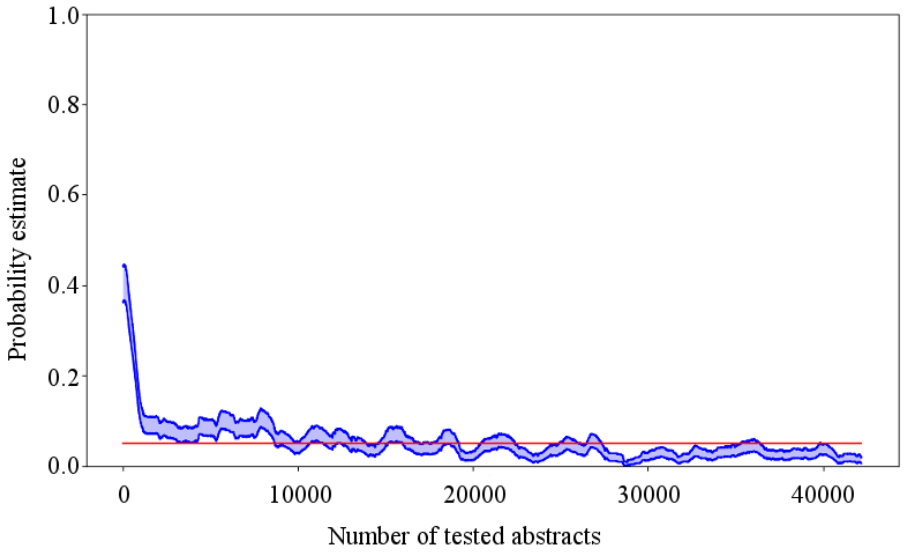

Fig. 4. 0.95 confidence interval of Poisson distribution of paper acceptance as a function of the number of already tested abstracts N and acceptance probability threshold 0.05.

The Citations Age To study the influence of a publication age on the probability of the publication citation we have attributed each edge of the PQA citation network with age calculated as difference between years of referencing publication and referenced one. The number of the edges as a function of edge age is shown on Figure 5. We can see that the maximal number of the references is observed for the publications that are 2–8 years old. Such publications are still regarded as new ones but at the same time are old enough to be read and estimated by many reserchers.

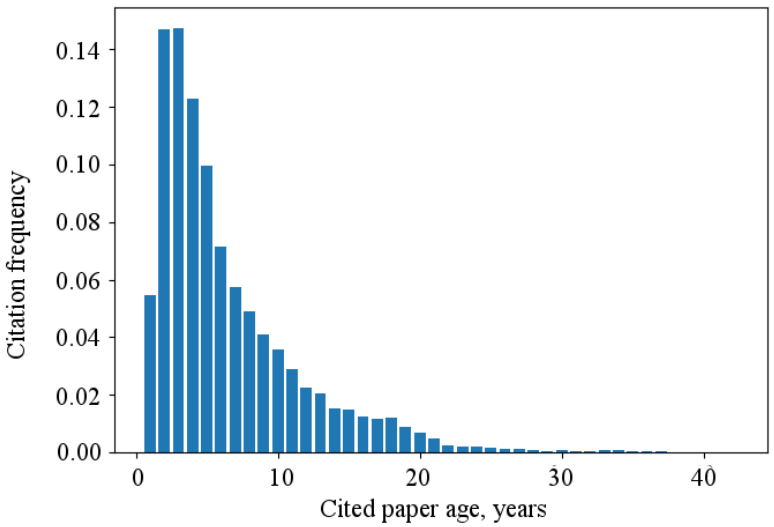

Fig. 5. Frequencies of the citation network edge ages.

The biased seed papers can produce unbiased citation network. To estimate the stability of the restricted snowball sampling with respect to the seed papers variation we run the sampling starting from the full set of the PQA seed papers and mark the relevant papers with main path analysis. Then we run the sampling again starting from 10 random subsets of the PQA seed papers and count the number of the runs where each seminal paper occurs. In our experiments the random subsets contain 50% of the seed papers and 66% of relevant papers are detected every time, 14% – in 80% of runs, 14% – in 60% of runs, 6% – at least once. So we can conclude that the PQA citation network is stable with respect to large seed paper variations being input for the restricted snowball sampling and the result of the sampling is unbiased.

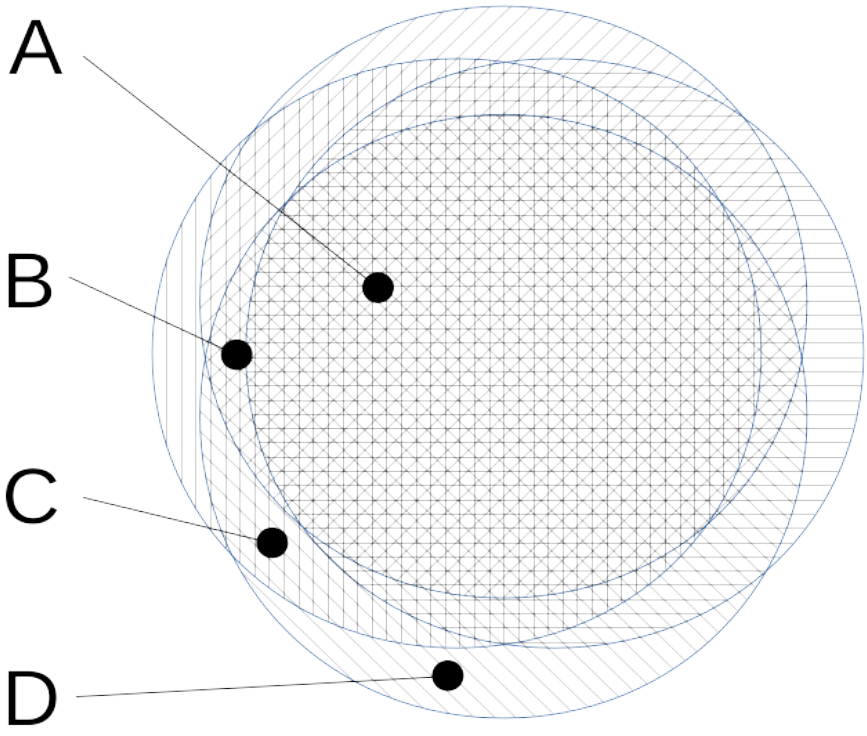

Fig. 6. Probability of the relevant paper detection: A – 66% detected every time; B – 14% detected in 80% of runs; C – 14% detected in 60% of runs; D – 6% detected at least once.

6 Conclusions and Future Studies

The main objective of the paper was to present the complete information technology that obtains a set of publications on some scientific topic as input and produces a list of seminal publications for that topic. It provides data for future detailed analysis and serves as a good point to begin investigation in a new domain. Additionally, we tested several initial assumptions regarding the results of the technology application and show that:

- PTM as a restriction criteria for the restricted snowball sampling method provides adequate semantic distance between publications.
- The restricted snowball sampling guarantees the saturation.
- The maximal number of the references is observed for the publications that are 2–8 years old. Such publications are still regarded as new ones but at the same time are old enough to be read and estimated by many reserchers.

– The biased seed papers produce unbiased citation network.
– The collected citation network is stable with respect to large seed paper variations being input for the restricted snowball sampling and the result of the sampling is unbiased.

The presented technology is implemented as sequence of Python scripts[3].

References

1. Ahad, A., Fayaz, M., Shah, A.S.: Navigation through citation network based on content similarity using cosine similarity algorithm. International Journal of Database Theory and Application 9(5), 9–20 (2016)
2. Barabási, A.L.: Scale-free networks: a decade and beyond. Science 325(5939), 412–413 (2009)
3. Batagelj, V.: Efficient algorithms for citation network analysis. arXiv preprint cs/0309023 (2003)
4. Batagelj, V., Mrvar, A.: Pajek-program for large network analysis. Connections 21(2), 47–57 (1998)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
6. Dobrovolskyi, H., Keberle, N., Todoriko, O.: Probabilistic topic modelling for controlled snowball sampling in citation network collection. In: International Conference on Knowledge Engineering and the Semantic Web. pp. 85–100. Springer (2017)
7. Ermolayev, V., Batsakis, S., Keberle, N., Tatarintseva, O., Antoniou, G.: Ontologies of time: Review and trends. International Journal of Computer Science & Applications 11(3) (2014)
8. Even, S.: Graph algorithms. Cambridge University Press (2011)
9. Golumbic, M.C.: Algorithmic graph theory and perfect graphs, vol. 57. Elsevier (2004)
10. Lecy, J.D., Beatty, K.E.: Representative literature reviews using constrained snowball sampling and citation network analysis (2012)
11. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge university press (2014)
12. Lucio-Arias, D., Leydesdorff, L.: Main-path analysis and path-dependent transitions in histcite-based historiograms. Journal of the Association for Information Science and Technology 59(12), 1948–1962 (2008)
13. MacKay, D.J.: Information theory, inference and learning algorithms. Cambridge university press (2003)
14. Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F.: A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics 61(1), 129–145 (2004)
15. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98(2), 404–409 (2001)
16. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences 101(suppl 1), 5200–5205 (2004)

[3] https://github.com/gendobr/snowball

17. Osborne, F., Motta, E.: Klink-2: integrating multiple web sources to generate semantic topic networks. In: International Semantic Web Conference. pp. 408–424. Springer (2015)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
19. Petticrew, M., Gilbody, S.: Planning and conducting systematic reviews. Health psychology in practice pp. 150–179 (2004)
20. Ráez, A.M., López, L.A.U., Steinberger, R.: Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In: Advances in Natural Language Processing, pp. 1–12. Springer (2004)
21. Salganik, M.J., Heckathorn, D.D.: Sampling and estimation in hidden populations using respondent-driven sampling. Sociological methodology 34(1), 193–240 (2004)
22. de Solla Price, D.J.: Networks of scientific papers. Science 149(3683), 510–515 (1965)
23. Valenzuela, M., Ha, V., Etzioni, O.: Identifying meaningful citations. In: AAAI Workshop: Scholarly Big Data (2015)
24. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: International Conference on Analysis of Images, Social Networks and Texts_x000D_. pp. 29–46. Springer (2014)
25. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1445–1456. ACM (2013)
26. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems 48(2), 379–398 (2016)