Probabilistic Topic Modelling for Controlled Snowball Sampling in Citation Network Collection

Hennadii Dobrovolskyi, Nataliya Keberle, Olga Todoriko

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

Abstract. The paper presents a probabilistic topic model (PTM) application to citation network collection. Snowball sampling method is moderated with the selection of the most relevant papers by means of the PTM. The PTM used in the paper is modified to treat collections of short texts. It is constructed from the titles of seed papers collection united with the papers obtained through unrestricted snowball sampling. The objective of the research is to propose and to experimentally verify the approach of application of PTM of short text documents for improvement of a citation network collection. The preliminary analysis has shown that the method is robust: seed paper collection variations do not affect the most influencing papers subset in the collected citation network.

Keywords: citation network, snowball sampling, text mining, short text document, topic modelling

1 Introduction

The completeness of a related work review is a problem well known to each scientist. The main question that should be answered is ”If the collected set of the scientific publications contain all significant knowledge of the domain of interest?”

The approaches designed to analyse an existing collection of scientific publications include citation analysis [1, 20], the study of the co-authorship [28, 29], elaboration of keywords and topics [27], examination of terminology [5]. The evaluation and adjustment of the mentioned methods are performed using different and thoughtfully prepared data sets [1] [2] [3], but the collection of a representative set of publications related to the particular topic is still the task of current interest. First, the existing collections of scientific papers do not cover

[1] AAAI Digital Library Conference Proceedings. https://aaai.org/Library/conferences-library.php

[2] Journals : Free Texts : Download & Streaming : Internet Archive. https://archive.org/details/journals

[3] Stanford Large Network Dataset Collection. https://snap.stanford.edu/data/

all the research topics and, second, gathering the publications manually is a time-consuming task which demands the expert involvement. In the actively evolving domains of research, manual search is obstructed with the fast growth of a number of publications. Another issue is the thesaurus diversity of different researchers which limit the completeness of the collected set of publication.

Individual variability of researchers knowledge and terminology affects the search because the authors of each scientific paper use their personal dictionary of terms which can be slightly different from terms used by the researcher looking for publications by keywords. As a result, the set of collected articles does not cover the whole domain of interest and the literature review is biased [32].

The objective of the presented paper is the implementation of restricted snowball sampling to build representative citation network of scientific publications on a domain of interest. To prevent infinite inflation of the sampled set we keep only the papers similar to the seed ones which are manually selected. The main question to be answered is if the sampling can collect most of the seminal publications concerning the domain of interest.

To estimate the publication similarity our algorithm uses probabilistic topic model. One of the challenges of similarity evaluation is that in most of the cases only titles and abstracts of the publications are available and the common methods like latent Dirichlet allocations [4] lose their precision. So the special modification combining word-word co-occurrence [40, 44], sparse symmetric nonnegative matrix factorization and principal component approximation [14] is suggested. It is demonstrated that the developed method of probabilistic topic modelling provides the natural estimation of the number of topics and allows calculation of the short text semantic similarity.

This paper is organized as follows. In section 2 we review existing advances in building citation network and estimating similarity of the short texts. Section 3 describes the used restricted sampling method along with a short discussion regarding the employed topic model, term dictionary reduction techniques, description of the suggested sparse symmetric nonnegative matrix factorization approach and the application of ideas of principal component approximation. In section 4 the experiment details are described and section 5 contains results of experiments. Section 6 is devoted to short conclusions and directions of the future studies.

2 Related Work

2.1 Building the Citation Network

Scientific search engines like Google Scholar[4], Microsoft Academic[5], Semantic Scholar[6] etc. provide different search facilities but do not answer the question

[4] Google Scholar https://scholar.google.com.ua/

[5] Microsoft Academic https://academic.microsoft.com/

[6] Semantic Scholar https://www.semanticscholar.org/

concerning the completeness of keyword search results with respect to a topic of interest.

One of the successful alternatives to keywords search is citation analysis [20] that creates and explores a directed graph which is a citation network. The nodes of the graph represent publications and the edge from node A to node B means that the publication A references publication B. References in a scientific article are selected thoroughly by the authors and advanced search engines can follow citations. It was shown [7] that citation analysis allows the creation of the more comprehensive list of publications than manual keyword search, it enables some formal description and smoothes individual differences of researchers.

The theoretical study of citation patterns was started by Price [34, 38] and Garfield [8]. It was demonstrated that detection of hubs simplifies the discovery of other parts of citation network [10, 15, 36]. The percentage of hubs which are the most cited papers is small because about 90% of publications are never cited [24].

The completeness of the citation network depends on the database coverage and on the quality of the search algorithm. In 2006 Meho and Kiduk [42] show that a single scientific search engine cannot provide enough data to collect the complete set of publications and several databases should be queried. But since then the scientific databases have been significantly extended and the coverage should be much better.

The high quality of citation-based search algorithm is ensured with phenomena of "small world" which is a proved property of scale-free networks [3, 28]. Newman [28] demonstrated that in the most of the cases it is enough to do three following iterations: each publication from the current queue is considered then all the papers it references and all papers referencing to the publication are added to the next level queue. The critical point of the algorithm is a selection of the initial queue which is called a seed collection. It should contain the papers that match the domain of interest and are widely cited.

The described iterations are the essence of the snowball sampling algorithm that is widely used to study social relations [17]. But it cannot be applied directly to publication crawling because the scientific texts often refer to the areas that are not directly related to the investigated domain. Therefore the straightforward implementation of snowball sampling causes infinite collection inflation [1] and some restrictions should be introduced while constructing the citation network. For instance, Ahad et.al in their approach uses [1] vector document model and cosine similarity to filter out the most relevant papers while sampling, however, when sampling the scientific abstracts the document vector model doesnt provide enough precision. Lecy et al [17] apply PageRank calculated by Google Scholar as a measure of paper significance. But PageRank is a property of a global citation network including all topics of knowledge, so it cannot be calculated from its small subset.

In this paper, we show that the restricted snowball sampling which utilises short text similarity can provide a set of publications which is small enough and contains most of the seminal papers concerning the domain of interest.

2.2 Similarity of Short Texts

In most of the scientific databases, the publication title, abstract, author names and the book or journal title are available, sometimes the database-specific keywords and topics are presented while full text is often protected by copyright. Therefore the only information we can rely on is the title and the abstract and we should decide if the paper matches the topic of interest where the topic is defined as a set of seed documents.

Natural approach suitable for such kind of matching is a short text similarity which can be calculated in many ways. String matching methods test whether words in two short texts are similar sequences of characters calculating largest common substring [12], edit distance [25] or lexical overlap [13]. However, string matching fails if the similar texts contain synonyms or have different word order.

Other methods of short text comparison are based on converting texts into syntactic trees [37]. A drawback of the syntactic parser is that it can process only texts having the well-defined structure and works only at the sentence level.

Semantic similarity calculation involves external sources of semantic knowledge such as WordNet [35]. However, such lexical databases are not available for all languages and even if they are, their dictionaries do not contain proper names, domain-specific terms, slang etc.

Recently developed deep learning approaches [26] only require a large amount of unlabeled data and represent words as vectors in a high-dimensional semantic space. Such a representation is referred to as embedding. The challenge is a transition from word level to sentence level [37]. While reducing the sequence of word vectors to a sentence vector of a fixed size some information is necessarily lost and it is crucial to decide which information to keep. Le and Mikolov suggested a variation of the word2vec algorithm for paragraph embedding [16]. Yang et al. [43] suggest an attention mechanism that maps the sequence of the word vectors to a single sentence vector. Also, the straightforward applications of neural networks to short text matching were developed [22]. Another method of word embedding, called GloVe, is proposed in [31]. It is based on word-word co-occurrence matrix and uses global matrix factorization, so it is close to BTM [40] and WNTM [44] statistical topic modelling. But all embedding methods require a lot of data that is not available in a restricted domain of knowledge. Moreover, the neural network embeddings are hard to interpret and adjust.

Distributional semantic approaches assume that words having similar context have the similar meaning. The famous LDA algorithm [4] uses this guess by building word-document co-occurrence matrix to get the probabilistic topic model (PTM). PTM allows mapping a text, sentence or separate word to a low-dimensional vector of topic probabilities. The distance between vectors can be used as the text similarity measure. However the critical drawback is the low precision when handling short texts. To overcome the last obstacle the word-word co-occurrence matrix [40, 41, 44] or other topic model modifications [30] can be used.

Below we use the word-word co-occurrence frequencies to build the probabilistic topic model and calculate similarity of scientific abstracts. We apply the sparse symmetric nonnegative matrix factorization together with the principal component analysis as a mean to naturally define the number of topics.

3 Restricted Snowball Sampling Method Description

The general workflow of the restricted snowball sampling contains the following steps:

1. Collect a set of seed papers;
2. Start from the seed papers and run several iterations of the unrestricted snowball sampling to gather baseline documents;
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 reduced dictionary containing all the significant words occurring in the sentences;
   4 combine all words 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 symmetric nonnegative matrix factorization (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:
   1 get a portion of papers from 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;Word-Word Co-Occurrence and Probabilistic Topic Model
   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.

Some of the listed steps requiring detailed explanation are discussed below.

3.1 Seed Paper Selection

Selecting the seminal publications is important for creating a comprehensive citation network. In [17] the authors recommend 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. Proper analysis of the dependence of the collected citation network recall on the seed publication properties is still an open question.

3.2 Word-Word Co-Occurrence and Probabilistic Topic Model

Let us consider a set of documents D and a dictionary W containing all terms used in D. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_{n_d})$ where "term" stands for a word or group of words. As well as common document topic model [39] the method used assumes that

1. Each term w in the document d is related to a topic t from a set of topics T. The document d is formed as a set of pairs (w, t), independently selected in a random way from discrete probability $p\,(d, w, t)$ defined over set $D \times W \times T$. The document d and the term w are observable and the topic t is the hidden parameter.
2. Order of terms in the document doesn't affect the topic model.
3. Order of documents in the collection doesn't affect the topic model.
4. Conditional probability $p\,(w|d, t)$ is independent on the document d, i.e. $p\,(w|d, t) = p\,(w|t)$.
5. The number of significant topics is far less than the number of words and the number of documents.

Like Biterm Topic Model [40] and Word Network Topic Model [44] the suggested method utilizes the joint probability $p\,(w_i, w_k)$ that both word w_i and word w_k occur in the same document or document fragment

$$p\,(w_i, w_k) = \sum_{t=1}^{T} p\,(w_i|t)\,p\,(t)\,p\,(w_k|t) \tag{1}$$

where t identifies a topic. The probability $p\,(w_i, w_k)$ is estimated as relative number of term pairs (w_i, w_k). To count the pairs each document d_k is mapped to the set of short sequences of terms $S\,(d_k) = (s_{k1}, s_{k2}, \ldots)$, where $s_{kq} = (w_{kq1}, \ldots, w_{kqr})$. Next, each sentence s_{kq} is mapped to pairs (w_i, w_k), $w_i \in s_{kq}$, $w_k \in s_{kq}$, $w_i \neq w_k$.

To build a topic model we need to evaluate the probabilities $p\,(w_i|t)$ and $p\,(t)$. Then detection of the document topics $p(t|d)$ is performed using the expression

$$p(t|d) = \sum_{i=1}^{|W|} p\,(t|w_i)\,p\,(w_i|d) \tag{2}$$

where $p(t|w_i)$ is found from the Bayes equation

$$p(t|w_i) = \frac{p(w_i|t)p(t)}{p(w_i)} \tag{3}$$

$p(w_i)$ is the probability of word w_i in the collection,

$$p(w_i) = \sum_{j=1}^{|W|} p\,(w_i, w_j) \tag{4}$$

and $p(w_i|d)$ is the relative frequency of word w_i in document d.

Because the documents are represented through their topic probabilities the difference between them can be measured with Kullback-Leibler divergence, Fisher's χ^2 or other statistical distances [23].

3.3 Dictionary Reduction

In the presented work, the size of joint probability matrix is reduced with stemming[7], keeping only nouns and adjectives, omitting stop-words and rare words.

Words which are not nouns and not adjectives can be excluded with part-of-speech tagger[8] because of small contribution to document topic assignment [33].

Stop-words are the terms that do not affect topic detection. Various lists of common stop-words are available online[9] but the collection-specific list has to be constructed.

To extract a set of collection-specific stop-words the probability $p\left(w_i, w_j\right)$ is employed. The background idea is that the stop-word w_i can co-occur with any other word so it has a large value of the Shannon information entropy

$$I(w_i) = \sum_{j=1}^{|W|} p\left(w_i, w_j\right)\ log\left[p\left(w_i, w_j\right)\right] \tag{5}$$

So the N_s terms w_i having the largest values of $I(w_i)$ are considered as stop-words and N_s can serve as the additional parameter of the algorithm.

Rare words are detected with the comparison of the single word probability $p(w_i)$ and a threshold value P_r. To define P_r we require that the kept terms cover $\alpha\%$ of occurrences.

$$\alpha = \sum_{p(w_i) \geq P_r} p(w_i) \tag{6}$$

We need to exclude the rare words from further consideration because of the statistical nature of our topic model. The small number of the rare words don't allow the reliable estimate of joint probabilities and keeping them we decrease the accuracy of the PTM.

After all the surplus words are excluded the joint probability matrix P becomes much smaller and should be decomposed into product of three matrixes according to (1).

3.4 Sparse Symmetric Nonnegative Matrix Factorization and Principal Component Analysis

The decomposition (1) can be simplified by defining matrix H such that

$$H_{it} = p\left(w_i|t\right)\sqrt{p\left(t\right)} \tag{7}$$

[7] See NLTK Stemmers http://www.nltk.org/howto/stem.html

[8] See NLTK part-of-speech tagger http://www.nltk.org/book/ch05.html

[9] For instance list of English stop words is available at Snowball stemmer site http://snowball.tartarus.org/algorithms/english/stop.txt

Dimensionality of the factor H_{ij} is $|W| \times |T|$, where $|W|$ is number of words in the reduced dictionary and $|T|$ is a suggested number of topics.

After the substitution (7) performed the Eq.(1) becomes well known Symmetric Nonnegative Matrix Factorization (SNMF) problem [9].

$$P \approx HH^T, H_{it} \geq 0 \tag{8}$$

Typically SNMF is formulated as optimization problem

$$\left\| P - HH^T \right\|_F^2 \to min, H_{it} \geq 0 \tag{9}$$

where $\|Z\|_F^2$ denotes squared Frobenius norm of a matrix Z.

Non-negative sparse coding [11] is a decomposition in which the factor H is sparse - it depends only on a few significant parameters improving human interpretability of the results. Usually [11] the sparsity is achieved with adding extra term to objective function (9):

$$\left\| P - HH^T \right\|_F^2 + \lambda \sum_{i,t} |H_{it}| \to min, H_{it} \geq 0 \tag{10}$$

where the parameter λ affects both sparsity level and factorization accuracy. The particularity of the Sparse Symmetric Nonnegative Matrix Factorization (SSNMF) defined with (10) is the simultaneous requirements of symmetricity, non-negativity and sparsity.

Similarly to other optimization problems (10) can be solved with projected gradient descent approach [19] which consists of the following update rule

$$H_{it}^{(n+1)} = max\left(0, H_{it}^{(n)} + \delta \nabla_{it}\right) \tag{11}$$

where ∇_{it} is a gradient of objective function (10) that takes into account symmetry of the matrix P:

$$\nabla_{it} = 4\left[\sum_j \left(\sum_p H_{ip}H_{jp} - P_{ij}\right) H_{jt}\right] + \lambda \tag{12}$$

and δ is a variable step size which is gradually decreased during iterations.

The main point of the presented method is that the number of topics T can be determined from the solution of (10). Below we demonstrate that sparsity condition leads to very small values of some topic probabilities. So their values can be set to zero and corresponding topics can be neglected in the resulting topic model. The used approach differs from the well known method of Principal Component Analysis [14] with the method of matrix decomposition.

After the factor H_{it} is calculated, the topic probabilities can be found as squared norm of matrix columns

$$p(t) = \left(\sum_{i=1}^{|W|} H_{it}\right)^2 \tag{13}$$

and the normalized columns are the topic word probabilities

$$p\left(w_i|t\right) = H_{it}/\sqrt{p(t)} \tag{14}$$

4 Experiment Details

In our experiments, seed papers are collected with searching for keywords ”automatic pronunciation assessment” in the Google Scholar database. The set of the seed papers is incomplete and imbalanced but it matches the main objective to test if we can collect the seminal papers starting from such a bad seed.

The current implementation of the restricted snowball sampling uses Academic Knowledge API [10] to search for scientific publication. The API allows to follow the citations in any direction and we can select both the papers referenced by current publication and the papers referencing the current publication. Also the API provided title, abstract, author names and list of topics.

To get the difference between two vectors of topic probabilities v_1 and v_2 we use the Kullback-Leibler divergence

$$D(v_1|v_2) = \sum_i v_1(i)\,\log\frac{v_1(i)}{v_2(i)}. \tag{15}$$

and distance to the set of seed papers is calculated as KL-divergence with the closest entry of the set. Then the distance is compared to a threshold to decide if the publication will be ignored or added to the snowball.

The parameters used in the experiment are: the upper bound of the topic number is set to 200; the percentage of stop words to exclude equals to 0.02; the percentage of rare words to exclude is 0.05; the sparsity parameter λ is 0.005; the threshold distance is 0.2; the number of seed papers is chosen to be 100 and random subset of seed papers has size 50; the number of levels of snowball sampling is 3; the number of runs is set to 12.

For our experiments, we selected a large set of seed papers from ”pronunciation quality assessment” domain and run the sampling method starting from random subsets of that set. Then the following average measures are calculated: the number of citation inside the collected network and the probability of detection with sampling. The number of citations is used instead of the total number of citations because the citation in the relevant publication is more valuable than the same citation in other domain of knowledge. The probability of a paper detection is a percentage of sampling runs discovered the paper. This measure shows how closely is the paper linked to other relevant papers.

[10] https://azure.microsoft.com/en-us/services/cognitive-services/academic-knowledge/

5 Results and Discussion

5.1 Stop-words Detection

Figure 2 shows dictionary terms ranked by values of information entropy calculated using the results of the run of unrestricted snowball sampling which is necessary to create the probabilistic topic model. The Shannon information entropy as function of the rank does not have some special points. It demonstrates the smoothly increasing dependency and we cannot point out some natural threshold. Thus the number of omitted stop-words remains an arbitrary external parameter of the developed algorithm allowing some adjustments.

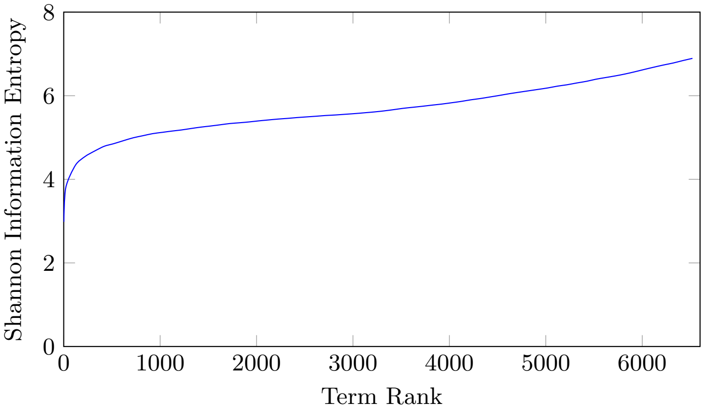

Fig. 1. Shannon information entropy as function of the term rank where the terms are ranked by value of the entropy.

The top 5 of collection-specific stop-words having the largest values of the entropy are ”article”, ”need”, ”effect”, ”main”, and ”aspect” which are a common terms specific to the scientific writing style. Those terms can be dropped out without loss of precision of the topic model.

5.2 SSNMF and PCA for Topic Modelling

In our experiments each of SSNMF runs starts from random initial state and in general case produces a different number of topics. However figure 3 demonstrates that in each case the number of topics can be determined in natural way because probabilities of some topics are too small to be significant and there is a critical topic rank where the topic probability sharply decreases to very small values.

However, the smaller number of topics does not mean the better model quality. Table 1 shows that when number of topics gets bigger the average topic coherence [2] increases i.e. the accuracy of the topic model increases.

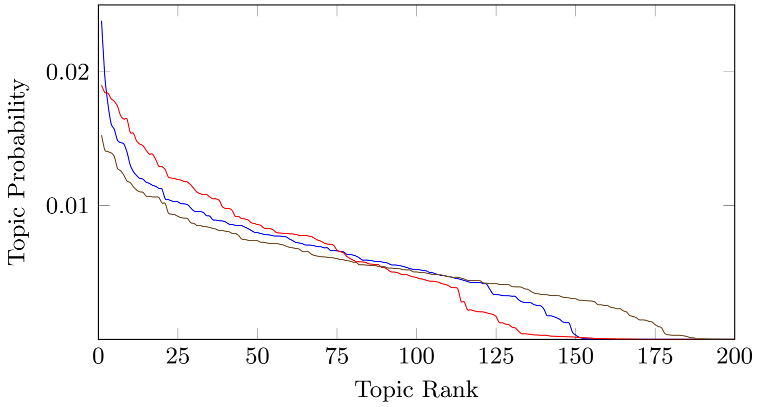

Fig. 2. Topic probabilities for different SSNMF runs for sparsity parameter $\lambda = 0.005$ and random initial states as function of topic rank. Topics are ranked by decrease of values of topic probability.

Table 1. Dependence of average topic coherence on number of topics

Snowball Run	Number of Topics	Average Coherence
1	135	0.471
2	151	0.481
3	180	0.487
4	182	0.519

5.3 Restricted Snowball Sampling

In the restricted snowball sampling, the mathemathically grounded recall evaluation is still an open question because the analogue of central limit theorem is not known for citation sub-networks as well as their statistical properties [17]. The most promising observed feature of the restricted snowball sampling related to the recall is the saturation. The figure 4 shows the probability that the publication will be accepted and added to the snowball. More detailed insight shows that the publications added at last stages of the sampling represent the related topics but do not directly concern the domain of interest. So we can conclude that the right part of the plot is generated with topic model inaccuracy rather than actual data. Hence we can stop sampling much earlier than we did.

It should be noted that new papers in the field will be added to the snowball as referencing already published relevant papers but they will not appear in the list of the most important articles because it takes time for others to read, understand and refer to such papers.

Next feature of interest is the level of precision. To measure the precision we extract top 50, top 100, top 150 etc. most cited papers, then mark the most relevant ones with hands and calculate their percentage as the function of detection probability and size of the most cited papers set. The table 2 shows that most

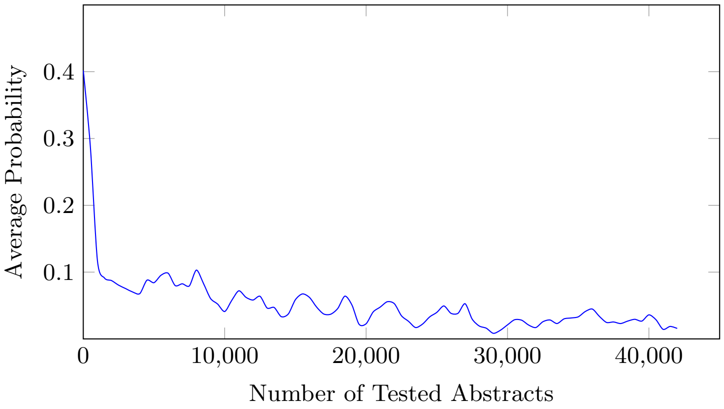

Fig. 3. Average probability of the situation that the publication will be added to the restricted snowball as a function depending on the number of tested abstracts.

of the relevant publications were sampled at each snowball run and multiple sampling runs slightly increase the method precision.

Table 2. The percentage of valid publications as function of the probability of paper detection where N_{top} is the number of the most cited papers extraced and P_{detect} is probability of paper detection

N_{top}	**Interval of P_{detect}**					
	0-100%	**0-20%**	**20-40%**	**40-60%**	**60-80%**	**80-100%**
50	0.50	0.04	0.02	0.02	0.06	0.36
100	0.39	0.02	0.01	0.04	0.05	0.27
150	0.36	0.02	0.01	0.05	0.05	0.25
200	0.29	0.01	0.01	0.04	0.04	0.19
250	0.27	0.02	0.01	0.04	0.03	0.17

The more effective way to increase the precision is the topic model justification which can provide the more precise estimation of semantic distance between the scientific abstracts.

5.4 Expert Selected Papers Evaluation

Another way to evaluate the restriction method is the test of publications selected by experts to check if the publications will be accepted or rejected. For the preliminary tests, we have chosen the publications of Ann Lee [18], Jonás Fouz-González [6], and Victoria López et al. [21]. The first and second papers contain reviews of the domain of interest and the last one describes one of machine learning topic which does not mention the automatic pronunciation assessment. Then for each of the publications we calculate the average distance between the list of references and the set of seed papers and obtain the following estimates: average distance from references selected by Jonás Fouz-González [6] to seed papers is 0.1199 ± 0.0136, the references of Ann Lee [18] are slightly furter (distance is 0.1465 ± 0.0267), and review of Victoria López et al. [21] concerning different topic has the largest distance 0.1744 ± 0.0186. Regarding the last paper as baseline corresponding to different area of knowledge we can see that its' references are clearly separated from relevant ones.

6 Conclusions and Future Studies

The main objective of the paper was to develop a method of collecting scientific publications on a domain of interest. The method relies only on the information contained in most of databases, namely paper titles, abstracts, author names and, sometimes, keywords. It provides data for future analysis, including mainstream outline, detection of cutting edge ideas and emerging subfields.

In the presented study we propose the application of probabilistic topic model to restricted snowball sampling which is an intelligent crawling the paper citation network and collecting the papers similar to the seed collection. The starting point of the crawler is a collection of the seed papers found with keywords search in any of scientific search engine. The advantage of the proposed method is its independence on the external information like Google Scholar page rank or full text of publication which may be not available.

The used probabilistic topic model is adapted to handle collections of a short texts with a known approach utilizing the word-word co-occurence probability. The proposed modification of the topic model uses sparse symmetric nonnegative matrix factorization and provides a natural way to determine the number of topics which is similar to principal component analysis.

The experiments show that the snowball sampling demonstrates saturation and is tolerant to the collection of the seed papers where both features are restricted with topic model inaccuracy. So the main direction of future works is increasing the precision of the topic model with parameter justification or running multiple iterations. A complete analysis of the collected citation network using known methods like PageRank or Search Path Analysis is also planned.

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. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: IWCS. vol. 13, pp. 13–22 (2013)
3. Barabási, A.L.: Scale-free networks: a decade and beyond. science 325(5939), 412–413 (2009)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)

5. 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)
6. Fouz-González, J.: Trends and directions in computer-assisted pronunciation training. In: Investigating English Pronunciation, pp. 314–342. Springer (2015)
7. Garfield, E.: From computational linguistics to algorithmic historiography. In: Symposium in Honor of Casimir Borkowski at the University of Pittsburgh School of Information Sciences (2001)
8. Garfield, E., Merton, R.K.: Citation indexing: Its theory and application in science, technology, and humanities, vol. 8. Wiley New York (1979)
9. Gillis, N.: Introduction to nonnegative matrix factorization. arXiv preprint arXiv:1703.00663 (2017)
10. Harris, J.K., Beatty, K.E., Lecy, J.D., Cyr, J.M., Shapiro, R.M.: Mapping the multidisciplinary field of public health services and systems research. American journal of preventive medicine 41(1), 105–111 (2011)
11. Hoyer, P.O.: Non-negative sparse coding. In: Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on. pp. 557–565. IEEE (2002)
12. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD) 2(2), 10 (2008)
13. Jijkoun, V., Rijke, M., et al.: Recognizing textual entailment using lexical similarity (2005)
14. Jolliffe, I.T.: Principal component analysis and factor analysis. In: Principal component analysis, pp. 115–128. Springer (1986)
15. Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., Komiyama, H.: Creating an academic landscape of sustainability science: an analysis of the citation network. Sustainability Science 2(2), 221 (2007)
16. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1188–1196 (2014)
17. Lecy, J.D., Beatty, K.E.: Representative literature reviews using constrained snowball sampling and citation network analysis (2012)
18. Lee, A., et al.: Language-independent methods for computer-assisted pronunciation training. Ph.D. thesis, Massachusetts Institute of Technology (2016)
19. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural computation 19(10), 2756–2779 (2007)
20. Liu, J.S., Lu, L.Y., Lu, W.M., Lin, B.J.: Data envelopment analysis 1978–2010: A citation-based literature survey. Omega 41(1), 3–15 (2013)
21. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113–141 (2013)
22. Lu, Z., Li, H.: A deep architecture for matching short texts. In: Advances in Neural Information Processing Systems. pp. 1367–1375 (2013)
23. MacKay, D.J.: Information theory, inference and learning algorithms. Cambridge university press (2003)
24. Meho, L.I.: The rise and rise of citation analysis. Physics World 20(1), 32 (2007)
25. Mihalcea, R., Corley, C., Strapparava, C., et al.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI. vol. 6, pp. 775–780 (2006)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

27. 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)
28. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98(2), 404–409 (2001)
29. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences 101(suppl 1), 5200–5205 (2004)
30. Pang, J., Li, X., Xie, H., Rao, Y.: Sbtm: Topic modeling over short texts. In: International Conference on Database Systems for Advanced Applications. pp. 43–56. Springer (2016)
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
32. Petticrew, M., Gilbody, S.: Planning and conducting systematic reviews. Health psychology in practice pp. 150–179 (2004)
33. Popova, S., Khodyrev, I., Egorov, A., Logvin, S., Gulyaev, S., Karpova, M., Mouromtsev, D.: Sci-search: Academic search and analysis system based on keyphrases. In: International Conference on Knowledge Engineering and the Semantic Web. pp. 281–288. Springer (2013)
34. Price, D.: Citation measures of hard science, soft science, technology, and non-science. communication among scientists and engineers. Nelson, CE, Pollack, DK Heath Lexington Books Massachusetts (1970)
35. Ramage, D., Rafferty, A.N., Manning, C.D.: Random walks for text semantic similarity. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing. pp. 23–31. Association for Computational Linguistics (2009)
36. Small, H.: Visualizing science by citation mapping. Journal of the Association for Information Science and Technology 50(9), 799 (1999)
37. Socher, R., Huang, E.H., Pennin, J., Manning, C.D., Ng, A.Y.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Advances in Neural Information Processing Systems. pp. 801–809 (2011)
38. de Solla Price, D.J.: Networks of scientific papers. Science 149(3683), 510–515 (1965)
39. 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)
40. 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)
41. Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y.: Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: Proceedings of the 2013 SIAM International Conference on Data Mining. pp. 749–757. SIAM (2013)
42. Yang, K., Meho, L.I.: Citation analysis: A comparison of google scholar, scopus, and web of science. Proceedings of the American Society for Information Science and Technology 43(1), 1–15 (2006), http://dx.doi.org/10.1002/meet.14504301185
43. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., Hovy, E.H.: Hierarchical attention networks for document classification. In: HLT-NAACL. pp. 1480–1489 (2016)
44. 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)