Obtaining the Minimal Terminologically Saturated Document Set with Controlled Snowball Sampling

Hennadii Dobrovolskyi[0000−0001−5742−104X], and Nataliya Keberle[0000−0001−7398−3464]

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

Abstract. Collecting the scientific papers to write the Related Work section, keeping up-to-date expertise in the topic of interest, or studying new scientific direction is the ill-defined information need that does not allow certainty about the completeness of search results. The controlled snowball method suggested by authors in the previous papers was extended with the objective criterion of the result completeness that allows stopping the search. The criterion is based on the assumption that the complete document set contains all terms describing the topic of interest. So, appending new document to the complete collection does not extend the list of terms. In the experiments, we compare our method of gathering the scientific papers describing the topic "Ontologies (computer science)" with other three common approaches: search by automatical detected topic in "Microsoft Academic" database, a keyword search in Google Scholar database, and query ACM digital library with author keywords. For each of the collected sets, the automatic term extraction was performed, and the size of the minimal saturated ordered document set is found. It was shown that terminological saturation is observed for the sets collected with controlled snowball method and with topic search in "Microsoft Academic" database. Moreover, the proposed controlled snowball provides the 10% smaller document set.

Keywords: terminological saturation, minimal saturated document set, citation network, controlled snowball sampling.

1 Introduction

The research of search behaviour of scientists [2] show that in addition to the related work review typical tasks are the research of new trends, the support of awareness, search for reviewers and/or colleagues for joint scientific projects. All the aforementioned tasks are characterized by low specificity, the high volume of results and, consequently, long search time. For example, a scientific search for the task of studying a new theory can last months or even years. Analysis of modern search engines showed the lack of tools to increase the specificity and reduce the volume of results[25].

The specificity of the search task is a characteristic of its definiteness. For example, a task with high specificity is to search for the meaning of a known word in the dictionary, and a search engine user can accurately say that the search is successful. The low specificity of the task, such as the study of a new theory, does not make it possible to state with certainty that the search is completed and does not need to be continued to refine the results. Therefore, having a stop search criterion is an important way of handling the low specificity. Increasing specificity can also be achieved through diversity – the ability of the information system to discover relevant documents that are significantly different from those already known to the user. For example, in the case of keyword searches, a high-diversity system should include relevant documents that do not contain the words listed in the search query or their synonyms in the search results.

Previously, the authors of this paper proposed the method of controlled snowball [9], in which low specificity is overcome by building and analyzing a citation network. The purpose of this work is to complement the method developed by the authors in previous works by the criterion of stopping the search, which will reduce the number of documents found while maintaining a sufficient level of completeness.

2 Related Work

Table 1. The suitability of detection and selection methods to solve scientists' search problems.

Method	Stop criterion	Diversity	Minimal volume detection
Systematic review [14, 28]	Depends on expert		
Keyword search [31]	–	–	–
Content filtration method [29]	–	–	–
Systems of collaborative filtering[32]	–	–	–
Neighbor-based recommendations [30]	–	–	–
Graph-based recommendations [20]	–	+	–
Citation network analysis, Ahad et.al [1]	–	+	–
Citation network analysis, Lecy et.al [21]	–	+	–

The method of systematic review has a stop search criterion as well as results completeness criteria. In [28] it is proposed to stop at the moment when a researcher understands that incorporation of new publications does not influence the conclusions made. The focus of concepts and not on publications [14] allows selection of the most important and helps to decide how to group and to analyze the selected publications. The disadvantage of the systematic review method is

its informality and lack of automation – such methods do not offer automatic search and numerical quality measures.

Keyword search [31] is provided by well-developed search tools, but it has been shown that the keyword set is often inaccurate or/and incomplete [28]. To improve keywords, Petticrew and Gilbody [28] recommend to interview researchers working in a chosen field of study; if the interview cannot be conducted, it is recommended that a researcher [16, 8] examine the documents found carefully and change the set of key ones based on that knowledge. Another disadvantage of keyword search is the low variety of search results – the search engine does not include relevant documents in the search results that do not contain the keywords or their synonyms specified in the query.

Insufficient variety of search results is attempted to be overcome by recommender systems [6]: content filtration methods, collaborative filtering, neighbour- and graph-based recommendations.

Content-based filtering (CBF) systems [29] offer the user documents similar to those that the user has already viewed, but they have low diversity and ignore the quality and popularity of documents[12].

Collaborative filtering (CF) is based on the assumption that the user will find useful documents that similar users select [32]. The recommendations obtained are varied because they are based not on the similarity of the documents but the similarity of the preferences. [27]. However, the collaborative filtering of scientific publications is complicated by their large number compared to the number of readers [35], which does not allow reliable statistical estimates.

Neighbourhood recommendations include documents that are often found alongside some specified documents [30]. The advantage of such recommendations is to concentrate on relationships instead of similarities. Neighbourhood recommendations offer related but inconsistent documents and thus approach collaborative filtering.

Graph-based recommendation systems use existing links or assume their existence and build. For example, a citation network is a graph in which document nodes are connected by directed citation relationships [3]. Depending on the modeling objects edges are considered as citations [3, 20], relationship ¡¡published in¿¿ [3, 20, 39], authorship [3, 39]. Some authors build graphs creating artificial links [39]. To identify the most relevant recommendations, the numerical properties of the nodes are calculated on the constructed graph. Most often, a random walk is used to search for popular objects starting with one or more random nodes [20].

Building of a citation network with a snowball method and its analysis [34, 22, 21, 1] is close to graph-based recommender systems. The essence of the approach lies in the creation and analysis of a directed graph – citation network, where nodes are scientific publications, and an edge linking a node A with a node B means that A references to B. The advantage of the approach is that references in each publication are carefully selected by authors, The disadvantage of a list of references is its incompleteness and systematic bias. Due to the restrictions on publication size, authors have to provide only a general and lim-

ited description of the publications most relevant to their research [14]. It was shown [17] that citation analysis allows to create more complete publication sets than keyword-based search, makes formal description possible, and also smooths out the individual weaknesses of the researcher.

High search speed is ensured by the presence of hubs [18] – most cited publications. Their number is small because about 90% of scientific publications are never cited [24]. Additionally, high search speed and search completeness are ensured by a "small world" property, that is a proven property of citation networks[4]. That is why an average length of a path between any two random nodes is much less than the whole network size. Simulation of P2P networks of a similar structure shows that [26] in most cases it is enough to perform 2-3 iterations of controlled snowball [1, 21, 9]: for each publication from a current queue all the documents referenced and belonging to the selected topic are added to the next level queue. To select the documents for a given topic Ahad with colleagues [1] use vector document model and cosine similarity measure, Lecy et al. [21] used PageRank from Google Scholar to select important publications. In the previous work of the authors [9] probabilistic topic model was used.

3 The Method of Collection Gathering

The goal of the presented method is to retrieve from all available publications $\mathbb{D}$ the subset $\mathbb{B} \subseteq \mathbb{D}$, that contains elements matching the users information need, where the information need is an informal and sometimes implicit set of requirements to search results [31] and the **user** stands for a person that performs one of the scientific search activities [2, 25]. Following common practice [31], the publication is considered as relevant if it, from the users point of view, matches the users information need.

The method used in the presented study is based on several assumptions. *Assumption 1.* Information need consists of several informal requirements:

1. all the publications from the $\mathbb{B}$ belong to a given subject area [2];
2. all the publications from the $\mathbb{B}$ are important in a given subject area [2];
3. the size of the $\mathbb{B}$ that allows detailed study in an acceptable time [2];
4. the presence in the $\mathbb{B}$ of all the important terms of a subject area [13].

Assumption 2. In what follows we assume that an information need is partially represented with a set of publications each of which is related to the given subject area [38].

Assumption 3. Below we assume that due to low specificity of the information need [2, 25] the user may know some keywords from the subject area and can select the relevant publications, but does not have sufficient qualification to evaluate their importance and completeness of the collected publication set [2].

Assumption 4. Each publication $d \in \mathbb{D}$ can be mapped to a set of sentences $\mathbb{S}(d)$, and each sentence$s \in \mathbb{S}(d)$ – into a set of collocations $\mathbb{C}(s)$, that is a subset of all collocations $\mathbb{C}$, that can be found in $\mathbb{D}$, where collocation c is a word or a tuple of words, sentence s – is an ordered set of collocations, document – is

an ordered set of sentences, and publication is a structure consisting of texts, key words, metainformation and references list. Also, the term τ is a collocation labelling a concept in a given subject area.

Definition 1. [33, 5] Citation mapping is defined over a set of publications $\mathbb{D}$ as

$$REF : \{v\} \to \{u \in \mathbb{D} \,|\, v \, cites \, u\}, \; v \in \mathbb{D}. \tag{1}$$

By applying citation mapping to a certain publication u, one can obtain a set of referenced publications. By applying inverse citation mapping

$$REF^{-1} : \{u\} \to \{v \in \mathbb{D} \,|\, v \, cites \, u\}, \; u \in \mathbb{D}, \tag{2}$$

to a certain publication u, one can obtain a set of publications referencing to u.

Repeated citation mapping REF^k is defined as a result of multiple application of REF.

The mapping (1) defines a directed graph – citation network [33, 5]

$$N = (\mathbb{D}, \mathbb{E}) \tag{3}$$

with edges $\mathbb{E} = \{vu, \forall v \in \mathbb{D}, u \in REF(\{v\})\}$ and nodes $d \in \mathbb{D}$.

Assumption 5. [5] Citation network (3) is almost acyclic:

$$\left|\left\{d \in \mathbb{D} \,|\, \exists k \in \mathbb{N}, d \in REF^k(\{d\})\right\}\right| \ll \left|\left\{d \in \mathbb{D} \,|\, \forall k \in \mathbb{N}, d \notin REF^k(\{d\})\right\}\right|, \tag{4}$$

where $k \in \mathbb{N}$ – is a path length in a citation network.

Assumption 6. [19] The necessary condition for presence in a $\mathbb{B}$ all the important terms of a given subject area is terminological saturation of an ordered set of publications.

Assumption 7. Full text of a publication is not available for automatic access. Often copyright restrictions make it difficult to automatically access the full text of a publication. For example, search system Scopus requires a registration, taking several steps with the usage of e-mail, and buying access to a publication – the operation that may not be automatized. That is why in the proposed information technology full texts of publications are used at the very last steps when the list of selected publications is minimal.

Formal hybrid mathematical model of the process of bibliographic detection and selection – is a tuple

$$\mathbb{M} = \langle \mathbb{D}, REF, PTM, DocDiff, \delta, Snowball, \mathbb{B}_0, DocListDiff, \omega, SPC, MaxRank, Terms, Cvalue, thd \rangle, \tag{5}$$

- $\mathbb{D}$ - publications available for analysis;
- REF - citation mapping;
- PTM - presentation of the content of the publication;
- $DocDiff$ - publication difference measure;
- δ - marginal difference in publications;

– *Snowball* - snowball iteration mapping;
– $\mathbb{B}_0$ - snowball iteration starting point;
– *SPC* - publication weigth in a subject area;
– $DocListDiff$ - closeness measure of ordered sets of publications;
– ω - marginal closeness measure of ordered sets of publications;
– *MaxRank* - maximal rank of publication;
– *Terms* - mapping of $\mathbb{D}$ into set of terms $\mathbb{T}$;
– $Cvalue(\tau)$ - term weigth τ;
– *thd* - difference measure of term sets.

A subject area description in a model is defined with a set of seed publications $\mathbb{B}_0$ ($\mathbb{B}_0 \subseteq \mathbb{D}$, $|\mathbb{B}_0| \sim O(10)$), which at the same time is a starting point of snowball iterations. Seed publications should obey such conditions:

– publication theme is relevant;
– publication age - 2-14 years;
– publication is often cited in *relevant publications*.

It is important to note that the last item differs from typical recommendations [34] on how to select the seed publications for snowball, providing a better start for snowball iterations, however requiring more efforts from a user.

Document relevance to the subject area is calculated with the help of a probabilistic topic model of text documents (PTM) [37, 40, 36]. PTM presents a content of each publication $d \in \mathbb{D}$ as conditional probabilities

$$p(t|d) = PTM\,(d)\,, \tag{6}$$

showing probabilities of belonging of publication d to a topic t. Each topic t is defined with probabilities $p(\tau_i|t)$ of belonging of collocation τ_i to the topic t, and a-priori probability $p(t)$. In the presented model an modified PTM is used, which is based on restoring distributions $p(\tau_i|t)$ and $p(t)$ from collocations co-occurrence frequencies

$$p(\tau_i, \tau_k) = \sum_t p(\tau_i|t)p(t)p(\tau_k|t), \tag{7}$$

which is calculated by counting the sentences s, where both τ_i and τ_k are found.

Mapping publications to conditional probabilities allows the application of the statistical measures [7] to calculate the difference $DocDiff$ between publications.

In our experiment, we use Kullbach-Leibler divergence and its threshold δ that is chosen to keep the top 30% of the relevant publications during the first controlled snowball iterations.

Snowball iteration mapping is defined as:

$$\begin{aligned}\mathbb{B}_{i+1} &= Snowball\,(\mathbb{B}_i) \\ &= \bigcup_{v \in \mathbb{B}_i} \{v\} \cup REF(\{v\}) \cup REF^{-1}(\{v\})\big|_{DocDiff(v,\mathbb{B}_0)<\delta}\,,\end{aligned} \tag{8}$$

where $\mathbb{B}_i \subseteq \mathbb{D}$. The equation (8) differs from others [1, 21] (i) by the usage of topic model of text documents for calculation of difference between publications and (ii) by traversing the citation graph both in the direction provided by references and in the inverse direction.

Publication weight in the subject area

$$SPC_i : v \ \rightarrow \ \mathbb{N}, v \in \ \mathbb{B}_i, \tag{9}$$

is defined as search path count (SPC) measure [23, 5] calculated in subgraph $N_i \in N$ citation network (3), built on the edges $\mathbb{E} = \{vu, \forall v \in \mathbb{B}_i, u \in \mathbb{B}_i \cap REF(\{v\})\}$ and nodes $d \in \mathbb{B}_i$ after transformation of cycles into acyclic fragments using preprint transformation [23, 5].

SPC_i allows to find a rank $Rank_i(v)$ of each publication and define an ordered publication set we look for:

$$\begin{aligned} \mathbb{L}_i\left(MaxRank\right) = (v_k)_{k=1}^{|\mathbb{B}_i|}, & Rank_i(v_k) < MaxRank, \\ & Rank_i(v_k) \leq Rank_i(v_{k+1}), \end{aligned} \tag{10}$$

where maximal publication rank $MaxRank$ restricts a number of items in a ordered publication set and is defined by the requirement of fixed point of iterations(8) achievement and terminological saturation.

Within the framework of the developed model, the degree of closeness of ordered sets of publications $DocListDiff$ is calculated with Spearman rank correlation $\rho(\mathbb{L}_i, \mathbb{L}_{i+1})$, and the fixed point of iterations (8) is

$$|\rho(\mathbb{L}_i, \mathbb{L}_{i+1}) - 1| < \omega, i > i_0, \tag{11}$$

where ω – marginal closeness measure of ordered sets of publications, (10) is a parameter setting a level of variability of ordered publications set.

Terminological saturation of ordered publications set is defined with the following condition: adding Δ publication into the end of the list (10) leaves the term list almost unchanged.

$$\frac{thd(\mathbb{T}_i(MaxRank), \mathbb{T}_i(MaxRank + \Delta))}{\epsilon_i} < 1, \Delta > 0. \tag{12}$$

Mapping of publications $\mathbb{L}_i$ into set of terms $\mathbb{T}_i$

$$\mathbb{T}_i = Terms(\mathbb{L}_i) \tag{13}$$

is conducted by application to the combined text of publications a procedure of automatic term extraction, proposed in K. Frantzi, S. Ananiadou H. Mima [15] and improved in V. Ermolayev et al., [19] that defines a term weigth $Cvalue_i(\tau)$ in a publication set $\mathbb{L}_i(MaxRank)$, marginal value ϵ_i of term weigth and the measure of terms sets difference [13].

$$thd(\mathbb{T}_i, \mathbb{T}_j) = \sum_{\tau \in \mathbb{T}_i \cap \mathbb{T}_j} |Cvalue_i(\tau) - Cvalue_j(\tau)| + \sum_{\tau \in \mathbb{T}_i - \mathbb{T}_j} |Cvalue_i(\tau)| \tag{14}$$

Minimal terminologically saturated publication set if described with the equation (10), where

$$MaxRank = min\left\{ M \left| \frac{thd(\mathbb{T}_i(M), \mathbb{T}_i(M + \Delta))}{\epsilon_i} < 1 \right. \right\} \tag{15}$$

The overall model quality measure (5) is a number of publications $|\mathbb{L}_i|$ in the final ordered publication set, restricted with (10), (11), (12) and (15).

Figure 1 shows the general workflow of the controlled snowball implementation as UML activity diagram.

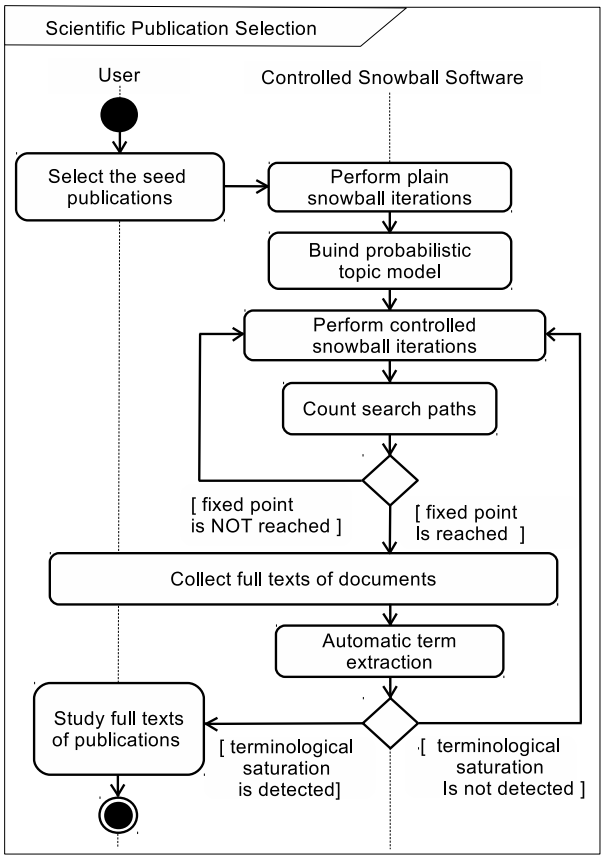

Fig. 1. General workflow of the controlled snowball implementation as UML activity diagram.

The general workflow was introduced in [10] and details of the restricted snowball sampling and probabilistic topic model construction are discussed in [11].

4 Terminological saturation of the ordered publication set obtained with controlled snowball method

The Spearman's rank correlation coefficient mentioned above allows simple detection of the convergence of controlled snowball iterations[10], however it does not address the completeness of the collected publication set.

The main idea of the presented experiment is comparison of minimal terminologically saturated ordered publication sets produced with different search methods and in different scientific databases and answer the following questions:

1. Do all common search methods produce the terminologically saturated ordered publication sets?
2. Which of the common search methods produce the smaller terminologically saturated ordered publication set?
3. Is the suggested controlled snowball method more effective than selection by topic?

In our experiments we concentrated on the existence of the terminological saturation and the size of the minimal terminologically saturated ordered document set that is defined as minimal value of *MaxRank* when (12) becomes true.

Starting from the uncertain information need of seminal scientific publications on the topic "Ontologies (computer science)", four collections were considered:

1. abstracts of the seminal publications selected from the ONTO-KL citation network that was gathered from the "Microsoft Academic Search" database using the controlled snowball method described above, starting from seed publications on the topic "Ontologies (computer science)";
2. abstracts of the publications indexed by the "Microsoft Academic Search" service having an automatically assigned category "ontologies" and arranged in descending order of citation index;
3. abstracts of the publications stored in the "ACM digital library" electronic library, having the "ontologies" label assigned by the authors and lined up in descending order of citation index;
4. abstracts of the publications found on Google Scholar Search by keyword "ontologies" and ranked by descending relevance calculated with Google's internal algorithms.

The 2nd, 3rd and 4th collections represent the common and wide spread search approaches that do not provide the formal criteria of stopping the search and thus can produce huge sets of publications. To enable comparison we have extended them with automatic term extraction and with method of terminological saturation detection.

For each of the found publications we searched for full text in PDF format. The PDF files were downloaded from different sources: "ACM digital library" provides full publication texts to registered users; "Microsoft Academic Search"

and "Google Scholar" often provide links to full-text PDF publications that can be automatically found and saved [1]. Also the PDF files were searched in SemanticScholar and ResearchGate databases. Publications for which the full text was not found were excluded from consideration and the text of the next publication was searched.

To study terminological saturation of ordered document set D we follow the work of Kosa et al. [19]. First, the finite sequence of texts D_i, $(i = 1, 2, 4, ..., 11)$ is composed where each text D_i contains the concantenated full texts of the first $20 \cdot i$ documents of D. Then all D_i are processed with the automatic term extraction method. The corresponding sets of terms T_i were compared with thd, defined by (14). The saturation criterion used is $thd(T_i, T_{i+1})/\epsilon < 1$ where $i \geq MaxRank$. Thus we can calculate minimal $MaxRank$ for any of used collections of publications.

The obtained values of minimal $MaxRank$ shown in the Table 2 are the quality measure for the proposed model. The Figure 2 shows the dependence of $thd(T_i, T_{i+1})/\epsilon$ from a number of publications incuded in D_i.

We can see that terminological saturation is observed for the collection gathered with the controlled snowball and selected from "Microsoft Academic". Publications gathered from "ACM digital library" do not provide saturation and set of publications taken from "Google Scholar" may exhibit saturation when extended.

The Table 2 shows that used in the paper controlled snowball method leads to smaller terminologically saturated publications set than studied analogues.

Table 2. The size of minimal terminologically saturated ordered sets of publications, belonging to topic "ontologies", obtained with different methods.

Source	**Search method**	*MaxRank*
"Microsoft Academic"	"Snowball"	160
"Microsoft Academic"	automatic label	180
"ACM digital library"	author label	> 200
"Google Scholar"	kew word	≥ 220

5 Conclusions

The presented study introduces the formal criterion of stopping the search and search result completeness to overcome the common issue of scientific information retrieval when information need has low certainty. The suggested formal criterion is based on automatic term extraction and terminological saturation detection.

[1] Software library Puppeteer for NodeJS, https://github.com/GoogleChrome/puppeteer

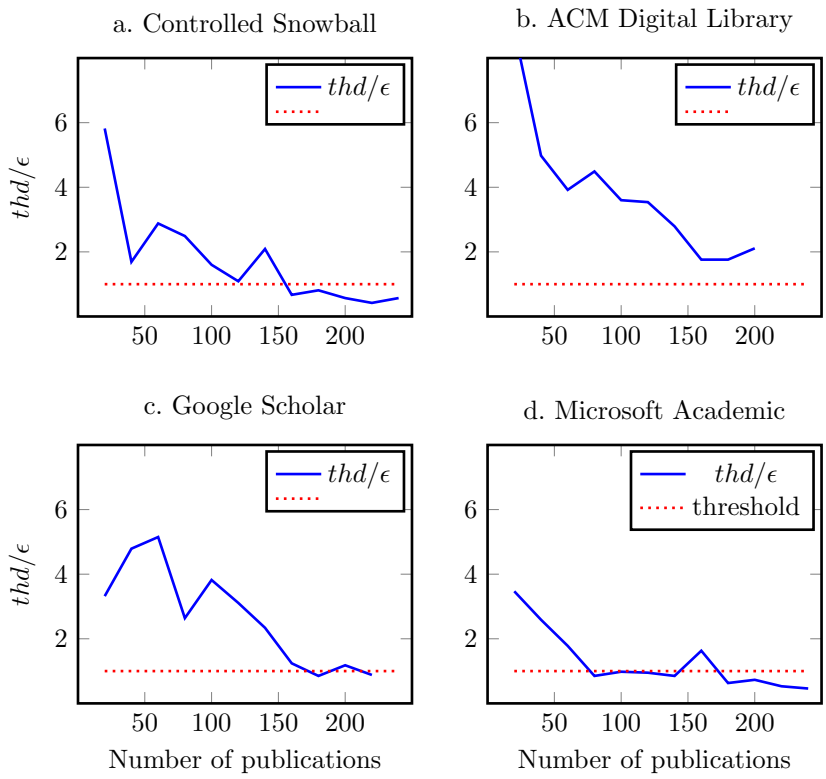

Fig. 2. Terminological saturation of publications collections.

In our experiment we have extended with terminological saturation detection the following search approaches: controlled snowball method; search by automatically assigned topic; keyword search; search by author keywords.

The objectives of the experiment were the existence of the terminological saturation and the size of the minimal terminologically saturated ordered document set for each of the search approaches.

Starting from the uncertain information need of seminal scientific publications on the topic "Ontologies (computer science)", four collections were considered:

1. publications gathered from the "Microsoft Academic Search" database using the controlled snowball method suggested by authors;
2. publications indexed by the "Microsoft Academic Search" service having an automatically assigned category "ontologies" and arranged in descending order of citation index;
3. publications stored in the "ACM digital library" electronic library, having the "ontologies" label assigned by the authors and lined up in descending order of citation index;
4. publications found on Google Scholar Search by keyword "ontologies" and ranked by descending relevance calculated with Google's internal algorithms.

The experiment have shown that terminological saturation for a collected ordered publication set, created from "Microsoft Academic" with the controlled snowball method, is achieved for 160 publications – 9% faster than for the ordered publication set created from "Microsoft Academic" with category "ontology" automatically set (180 publications). Sets consisting of 200 publications with the keyword "ontology" from "Google Scholar" and with label "ontology" from "ACM digital library", do not possess terminological saturation.

So we can conclude that both the controlled snowball method and topic search in "Microsoft Academic" produce the small terminologically saturated publication sets of almost equal size. However, this conclusion must be supported with search on other topics. Also, in the future studies, the term-based precision and recall should be calculated that, in turn, requires the creation of the dataset of terms evaluated by experts.

References

1. Ahad, A., Fayaz, M., Shah, A.S.: Navigation through citation network based on content similarity using cosine similarity algorithm. Int. J. Database Theory Appl **9**(5), 9–20 (2016). https://doi.org/10.14257/ijdta.2016.9.5.02
2. Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., Jacucci, G.: Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In: Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries. p. 20. American Society for Information Science (2013). https://doi.org/10.1002/meet.14505001041
3. Baez, M., Mirylenka, D., Parra, C.: Understanding and supporting search for scholarly knowledge. Proceeding of the 7th European Computer Science Summit pp. 1–8 (2011)

4. Barabási, A.L.: Scale-free networks: a decade and beyond. Science **325**(5939), 412–413 (2009). https://doi.org/10.1126/science.1173299
5. Batagelj, V.: Efficient algorithms for citation network analysis. arXiv preprint cs/0309023 (2003)
6. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Paper recommender systems: a literature survey. International Journal on Digital Libraries **17**(4), 305–338 (2016). https://doi.org/10.1007/s00799-015-0156-0
7. Choi, S.S., Cha, S.H., Tappert, C.C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics **8**(1), 43–48 (2010)
8. Colicchia, C., Strozzi, F.: Supply chain risk management: a new methodology for a systematic literature review. Supply Chain Management: An International Journal **17**(4), 403–418 (2012)
9. Dobrovolskyi, H., Keberle, N.: Collecting the seminal scientific abstracts with topic modelling, snowball sampling and citation analysis. In: Proceedings of the 14th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. vol. 1, pp. 179–192. Springer (2018)
10. Dobrovolskyi, H., Keberle, N.: On convergence of controlled snowball sampling for scientific abstracts collection. In: International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications. vol. 1007, pp. 18–42. Springer (2018). https://doi.org/10.1007/978-3-030-13929-2_2
11. 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). https://doi.org/10.1007/978-3-319-69548-8_7
12. Dong, R., Tokarchuk, L., Ma, A.: Digging friendship: paper recommendation in social network. In: Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009). pp. 21–28 (2009)
13. 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)
14. Fisch, C., Block, J.: Six tips for your (systematic) literature review in business and management research. Management Review Quarterly **68**(2), 103–106 (Apr 2018). https://doi.org/10.1007/s11301-018-0142-x, https://doi.org/10.1007/s11301-018-0142-x
15. Frantzi, K.T., Ananiadou, S.: The c-value/nc-value domain-independent method for multi-word term extraction. Journal of Natural Language Processing **6**(3), 145–179 (1999). https://doi.org/10.5715/jnlp.6.3_145
16. Friday, D., Ryan, S., Sridharan, R., Collins, D.: Collaborative risk management: a systematic literature review. International Journal of Physical Distribution & Logistics Management **48**(3), 231–253 (2018). https://doi.org/10.1108/IJPDLM-01-2017-0035
17. 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)
18. 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). https://doi.org/10.1016/j.amepre.2011.03.015

19. Kosa, V., Chaves-Fraga, D., Dobrovolskyi, H., Ermolayev, V.: Optimized term extraction method based on computing merged partial c-values. In: Ermolayev, V., Mallet, F., Yakovyna, V., Mayr, H., Spivakovsky, A. (eds.) Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2019. Communications in Computer and Information Science, vol. 1175, pp. 24–49. Springer Berlin Heidelberg (2020). https://doi.org/10.1007/978-3-030-39459-2_2, https://link.springer.com/chapter/10.1007/978-3-030-39459-2_2
20. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained random walks. Machine learning **81**(1), 53–67 (2010). https://doi.org/10.1007/s10994-010-5205-8
21. Lecy, J.D., Beatty, K.E.: Representative literature reviews using constrained snowball sampling and citation network analysis. Available at SSRN 1992601 (2012). https://doi.org/10.2139/ssrn.1992601
22. 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). https://doi.org/10.1016/j.omega.2010.12.006
23. Lucio-Arias, D., Leydesdorff, L.: Main-path analysis and path-dependent transitions in histcite-based historiograms. Journal of the American Society for Information Science and Technology **59**(12), 1948–1962 (2008)
24. Meho, L.I.: The rise and rise of citation analysis. Physics World **20**(1), 32 (2007). https://doi.org/10.1088/2058-7058/20/1/33
25. Nedumov, Y., Kuznetsov, S.: Exploratory search for scientific articles. Programming and Computer Software **45**(7), 405–416 (2019). https://doi.org/10.15514/ISPRAS-2018-30(6)-10
26. Nicolini, A.L., Lorenzetti, C.M., Maguitman, A.G., Chesñevar, C.I.: Intelligent algorithms for improving communication patterns in thematic p2p search. Information Processing & Management **53**(2), 388–404 (2017). https://doi.org/10.1016/j.ipm.2016.12.001
27. Palopoli, L., Rosaci, D., Sarné, G.M.: A multi-tiered recommender system architecture for supporting e-commerce. In: Intelligent Distributed Computing VI, pp. 71–81. Springer (2013). https://doi.org/10.1007/978-3-642-32524-3_10
28. Petticrew, M., Gilbody, S.: Planning and conducting systematic reviews. Health psychology in practice pp. 150–179 (2004)
29. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender systems handbook, pp. 1–35. Springer (2011). https://doi.org/10.1007/978-0-387-85820-3_1
30. Rodriguez-Prieto, O., Araujo, L., Martinez-Romo, J.: Discovering related scientific literature beyond semantic similarity: a new co-citation approach. Scientometrics **120**(1), 105–127 (2019). https://doi.org/10.1007/s11192-019-03125-9
31. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press (2008)
32. Shi, Y., Larson, M., Hanjalic, A.: Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Computing Surveys (CSUR) **47**(1), 3 (2014). https://doi.org/10.1145/2556270
33. de Solla Price, D.J.: Networks of scientific papers. Science **149**(3683), 510–515 (1965)
34. Varela, A.R., Pratt, M., Harris, J., Lecy, J., Salvo, D., Brownson, R.C., Hallal, P.C.: Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis. Preventive medicine **111**, 466–472 (2018). https://doi.org/10.1016/j.ypmed.2017.10.020

35. Vellino, A.: Usage-based vs. citation-based methods for recommending scholarly research articles. arXiv preprint arXiv:1303.7149 (2013)
36. 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). https://doi.org/10.1007/978-3-319-12580-0_3
37. 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). https://doi.org/10.1145/2488388.2488514
38. Zarrinkalam, F., Kahani, M.: Semcir: A citation recommendation system based on a novel semantic distance measure. Program **47**(1), 92–112 (2013). https://doi.org/10.1108/00330331311296320
39. Zhou, M., Zhao, S.: Learning question paraphrases from log data (Feb 14 2008), uS Patent App. 11/500,224
40. 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). https://doi.org/10.1007/s10115-015-0882-z