Leveraging Scientometric Citation Clusters for Improved Information Retrieval

Laxmi Vijayan

July 30, 2023

1 Specific Aims

Current search tools of academic literature most frequently rely on keyword-matching with minimal inclusion of citation data to identify and rank research articles relevant to a query. However, this method fails to consider a wealth of data presented by the citation relationships between two documents and strips the documents of contextual information. We posit that there is an opportunity to leverage scientometric methods of clustering based on citation relationships to improve existing information retrieval (IR) methods.

In Scientometrics, there is a body of work that aims to map the structure of science by clustering similar articles based on citation relationships. The most commonly studied citation relationships are direct citation, bibliographic coupling, and co-citation. Each describes different facets of the relationship between two documents: the first models information flow, the second examines static topic similarity, and the third serves as a proxy for dynamic topic similarity between two documents. There is a fourth, less-studied relationship, longitudinal coupling, which evaluates the similarity between two temporally distant documents. Taken in concert, these relationships can provide context and help determine the relevance of a document holistically. Consequently, citation clustering-based IR that incorporates this information may yield more relevant results.

The proposed research is the first phase of a multiphase project to develop and test various citation cluster-based strategies for improving information retrieval of academic literature. In this phase, we aim to build upon existing citation-based clustering methods to identify those that yield viable clusters, i.e., clusters with high recall and precision, for use within an IR context. In Scientometrics, several studies have previously evaluated the relative accuracy, defined as high topic similarity, of different citation relationships and clustering algorithms. Here, we would instead like to evaluate clusters based on a more broad definition of relevance by accounting for edge cases, such as methods papers that might be relevant to more than one topic. As described above, we assert that a network accounting for all four citation relationships

will yield the best result in terms of recall and precision. Thus, we will construct several graph partitions of different citation networks with the widely-used clustering algorithm, Leiden. Then, we will iteratively evaluate which partitions are likely to be most suitable for use within an IR system through the use of different internal and external criteria.

These results will be foundational for achieving our long-term goal: to develop citation clustering methods that can be utilized for effective information retrieval in academic literature.

2 Significance

IR systems represent the process by which information is stored, queried, retrieved, and presented. [30] Two parts, a database that stores documents and a retrieval process that conducts query matches, work in concert to present a user with a ranked set of relevant documents in response to a query. [15], [30] In traditional IR, queries are lexical in nature and the matching is based on textual similarity between the query and the stored documents. The inherent ambiguity of lexical queries in both, general and specialized contexts, is a major drawback of the traditional systems. [14] Another drawback is the significant time complexity associated with searching the full-texts of large datasets. [29]

One strategy that has been developed to address the time-complexity is document clustering-based IR. This strategy is based on the cluster hypothesis which states "closely associated documents tend to be relevant to the same requests" [26]. Several studies have documented an increase in efficiency due to document-based clustering, however, they have also noted a commensurate drop in effectiveness due to the challenges of identifying sufficiently accurate clusters based on textual similarity. [29]

A rich literature on clustering, referred to interchangeably as community detection, in citation networks from the field of Scientometrics documents the various flavors of networks and clustering techniques. Here, scientific publications are modeled as nodes and edges are represented by various citation relationships (i.e. direct citation, bibliographic coupling, and co-citation), each capturing different perspectives. [12], [19], [21] Networks have been clustered using a range of clustering algorithms, each of which emphasize and capture only an aspect of the complex definition of community. [6], [24] Communities, in this context, are loosely defined as publications more densely connected to each other, than to the network at large.

A mainstay of Scientometrics has been topic reconstruction. Several studies have examined the similarity between different citation relationships, different clustering algorithms, and the accuracy of the resulting clusters. [1], [5], [13] An underlying assumption of these studies has been that research topics are distinct and isolated, and can be accurately identified by disjointly partitioning scientific publications. Given the inherent ambiguity in the definition of community and the lack of any ground-truth, the validation of the results in these studies have relied on a variety of proxy measures. Since the relationship between research topics and these

proxy measures is unknown, this validation is, arguably, insufficient. [9] However, we believe their work has laid the foundation for how various citation relationships can be applied towards creating viable citation cluster-based IR strategies.

The development of citation cluster-based IR has several advantages over existing search systems. Citation clusters include a wealth of contextual information, such as information flow and topical similarity based on each document's respective bibliographies. One study examining a hierarchical citation cluster-based IR has shown that it results in high precision. [2] Citation cluster-based IR can mitigate the lexical ambiguity of traditional keyword-matching based search, and can go some way towards mitigating the drop in effectivity of document cluster-based IR, which relies on textual similarity. Given the large body of existing academic literature and the exponential rate of growth of scientific publications (i.e. 4% per annum) [4], this method retains the needed efficiency of document cluster-based IR.

The rapid growth rate of science also places an added burden on users of these search tools, in this case researchers and scientists. A cornerstone of any research project is identifying relevant literature by conducting comprehensive literature reviews. In the modern-day, these are usually conducted through the use of web-based literature search systems which have reduced search time. However, query resulting in large sets with low recall and high precision or vice versa that don't sufficiently capture the topology of the targeted research area will diminish any net gain in time and effort. [8] A citation cluster-based IR strategy has the potential to not only improve the quality of the results for users by placing publication within a greater scientific context, but also capture other facets of the citations that users use to determine relevance, such as author prestige. [17]

3 Innovation

As described previously, there are a range of studies that examine the similarities between different citation relations, including several that examine combined linkages. [1], [5], [13] These studies belong to one school of thought: they posit that there is some absolute degree of similarity between the different relationships. Another school of thought is that there is some degree of difference between relationships resulting from the use of different agents and viewpoints. [7] A lesser acknowledged school of thought is put forth by Von Luxburg, Williamson, and Guyon (2011) that suggests that different partitions should be evaluated based on their usefulness towards some final goal. [27] We ascribe to this school of thought and will not be comparing different partitions, but evaluating the quality of partitions independent of each other.

Cluster-based IR and citation based IR, specifically in academic literature, have been well-explored by the scientific community. [2], [3], [29] The former method usually examines clusters based on textual similarity, not citations, while the latter does not capture the structure of the whole citation network, but identifies the flow of information of one focal document by listing references for the document and citations that the document has received.

While citation cluster-based information retrieval was proposed and discussed several decades ago, there has been little accumulated research on the topic. [2], [23] The work most similar to ours is by J.P. Bascur et. al., (2023). We expand on their work by including different citation relationships and a flat clustering (as opposed to their hierarchical clustering). We also use different methods to evaluate the viability of partitions. We are not yet interested in measuring the capability of a citation clustering to effectively satisfy a search task. Instead, we are determining whether citation based-clustering can be calibrated to increase recall while maintaining the precision they attained in their study.

4 Approach

4.1 Data and Pre-Processing

Our study corpus needs to be sufficiently large to ensure accurate representations of different citation relationships. Further, we need an external classification system to verify the relative quality of our partitions. Therefore, we will be renting the Dimensions.AI database. This database contains upwards of 95 million publications, as well as over 990 million citation links. [10] As part of the data curation of the Dimensions database, a rigorous classification of articles has also been developed using machine learning. While open source databases exist, it is expected that the rental service will have a more comprehensive and accurate information.

Any publications will be excluded from our dataset if they:

- do not have any direct citation relationships with other publications within the dataset,
- are duplicates, missing DOI, or publication date information,

Pre-Processing: Publications with high in-degree (cited frequently) and publications with high out-degree (long bibliography) will be examined and tagged as either "method" or "review." This will allow us to treat them differently during the post-processing stage.

4.2 Networks

4.2.1 Graph Notation

We will be modeling community detection as a graph partition problem. Let $G = (V, E)$ be a graph where $n = |V|$ nodes, $m = |E|$ edges. For each edge, let w_{nm} be a non-negative weight greater than 0 that signifies the strength of the relationship between the two corresponding nodes. Our clustering will be non-exhaustive, so some nodes may not be included in any cluster, and we will be doing some post-processing of the clustering which may result in overlapping clusters.

4.2.2 Citation Relationships

Different citation networks can be constructed using different citation relations. As mentioned before, the three most common are direct citation relationships, bibliographic coupling, and co-citation relationships. A fourth less studied citation relationship was put forth by Small in 1995 which described an indirect relationship between documents: longitudinal coupling. [22] This relationship was further explored as part of a combination linkage to improve recall in clusters. [20] While this was not pursued in later publications, likely due to the computational complexity of calculating two-step connections over a large network at the time, we believe it may mitigate some of the effects of rapid attention decay in science and contribute to a more wholistic perspective. [18] We have decided to include a modified version of this coupling in our networks. The definitions of the different relationships and calculations associated with the different relationships are listed below:

4.2.3 Direct Citation (DC)

Direct citation relationships are first-order relationships that track flow of information within the citation network. A direction citation relationship between nodes n_i and n_j exist if i cites or is cited by j. For the purpose of calculating weight in other networks, this will be quantified as 1 when an edge exists and 0 otherwise.

4.2.4 Bibliographic Coupling (BC)

Bibliographic coupling relationships are second-order relationships that serve as proxy measure of similarity between two documents within a citation network. They are determined by set operations on the out-edges of two nodes n_i and n_j where if i and j have out-edges pointing to the same third entity, there exists a bibliographic coupling. The weight of this edge is determined the number of entities they have in common.

4.2.5 Co-citation (CC)

Co-citation relationships are second-order relationships that serve as proxy measure of the relevance of two articles to a child article. It can be interpreted as either similarity between the articles, or a recombination of disparate topics resulting in a new research domain. If two nodes n_i and n_j are cited together by third entity, then they have a co-citation relationship. The weight of this edge is determined by the frequency at which they are cited together.

4.2.6 Modified Longitudinal Coupling (mLC)

Longitudinal coupling relationships are first discussed by Small (1995) as a way to achieve a more accurate global mapping that accounted for time. We revisit this work with a minor modification to account for the computation complexity that would arise if these weights were calculated for every pair of publications within our dataset. If two nodes n_i and n_j have a

co-citation relationship determined by n_c and a bibliographic coupling determined by n_b, then there will exist an edge between n_b and n_c such that $w_{cb} = \frac{w_{CC_{ij}} * w_{BC_{ij}}}{2}$. We use a half-weight relationship to account for the two-step connection between c and b that is proportional the bibliographic coupling and co-citation relationship weights of n_i and n_j. This relationship will create cliques between closely related pairs, and we believe that it will minimize the impact of any weak inter-disciplinary connections of co-citation relationships.

4.2.7 DC-BC-CC-mLC

Lastly, we will create a network that is determined by the sum of the previously listed relationships. No weight will be given to any particular relationship.

Given the time complexity for conducting these calculations for every pair within our large dataset, they will only be conducted once.

4.3 Community Detection

We have chosen to utilize the Leiden algorithm due to its fast performance with large networks. [25] The default optimization criterion for the Leiden algorithm is the the Constant Pott's Model (CPM) quality function. This function uses a linear resolution parameter (γ). It can be written in several ways and one formulation is a summation of communities:

$$Q = \sum_c \left[m_c - \gamma \binom{n_c}{2} \right] \tag{1}$$

where m_c is the total number of edges inside the community c and $\binom{n_c}{2}$ represents the number of possible edges within the community (n_c represents the number of nodes in community c).[25] By summing the difference between the total number of edges and the total possible number of edges over all communities, the CPM function determines the difference between the actual number of edges and the expected number of edges in a random network. Maximizing this value helps identify more densely connected communities. The resolution parameter, γ, allows for adjusting the granularity of the clustering, where smaller values return smaller, more dense clusters. Leiden guarantees that the internal edge density of communities is higher than γ, while the external edge density of communities is lower. Consequently, we will calculate the average graph density of the unclustered network, round to the closest 10^{-n} order of magnitude for each network and increment resolutions until the order of magnitude of 10^{-2} for each network.

We will follow the clustering properties put forth by Šubelj, van Eck, and Waltman (2016) where ever possible:

- The largest clusters will be of a size no more than ten times the smallest cluster. This will exclude clusters too large or small to merit any meaningful interpretation.

- Any clusters beneath $\log_{10}(n)$, where n represents number of nodes within the cluster, will be removed. As above, too small clusters will not contribute meaningfully to the task.
- Different clusterings should have some threshold of stability. The threshold will be determined iteratively. Stability over clustering will serve as one proxy to ensure that the clustering is not arbitrarily determined.
- Computing time should be reduced where possible. A long-term goal of this study is to identify efficient clustering strategies, therefore minimizing the clustering time is a priority.
- Some basic ground-truth must be verifiable. This will be determined through the use of our marker nodes, as well as during our qualitative evaluation process.

4.3.1 Calibration and Post-Processing

Basic Ground-Truth: In order to verify basic ground truth, we will utilize a marker node strategy as described in Wedell et. al. (2022). [28] We will use the dataset they have made available on Github. [16] This dataset contains 1,218 articles cited in 12 recent reviews on extracellular vesicles and exosome biology. We will verify if the articles are a.) present in within the clustering solution and b.) confirm that there are clusters with higher concentrations of markers.

Calibration and Post-Processing: As discussed previously, we will cluster the network at several resolutions. These clusters will be evaluated using partition metrics such as modularity and conductance. In order to obtain these statistics, we will be using the Connectivity Modifier++ to filter out any singletons, clusters of size 1, and any small clusters determined by our threshold. [11] For each cluster determined by Leiden at a given resolution, cm++ will iteratively process clusters to remove any trees, stars, and break apart weakly-connected clusters. The pipeline also provides statistics for each cluster within a clustering before and after processing, as well as overall clustering statistics such as node count and node coverage that will allow us to evaluate the quality of the partition. Clustering that show low community structure or low node coverage will be excluded from any further analysis. In each clustering, we will identify the high degree papers that were tagged as either "review" or "method." For each review paper, we will verify if it has been included within a cluster and that 80% of it's references are also in the same cluster. For each "method" paper, we will assign it to each cluster that references it frequently. The threshold for including it within a cluster will be determined empirically during our initial clustering and calibration process.

4.4 Evaluation

Aside from the metrics described above, we will be using qualitative and quantitative evaluation to determine which resolution has maximizes topical similarity while maximizing cluster size.

4.4.1 Qualitative Evaluation

A small subset of random tuples of documents, (A, B, C), will be extracted. Two documents will be from the same cluster (A, B), and one document will be from another cluster (C). Document A will be a focal document, while B and C will be used for comparison. These documents will be matched to their title and abstract obtained from the Dimensions database. Finally, a questionnaire will be created to compare documents A and B and documents A and C. To collect human judgements of relevance, we will recruit crowdsource workers from a crowdsourcing platform. Different crowdsourcing platforms and their merits will be evaluated in order to ensure there are some built-in quality metrics. Recruited participants will be given some training on how to respond to the questionnaire, and the following questions will be presented to them:

- Which paper, B or C, do you think is most similar in topic to A?
- What keywords would you use to categorize both A and [B or C]?

Some quality control metrics that will be considered will be the amount of time each participant spends with reading the three titles and abstracts, as well as the use of random or unrelated keywords to describe the two articles. The sample size for number of document tuples and participant sample size will be determined based on further discussion with our statistician.

4.4.2 Quantitative Evaluation

We will also leverage the existence of categorization provided by the Dimensions database to evaluate the topical similarity of the clusters.

Intracluster similarity will be determined using Jaccard Similarity of the keywords for document pairs:

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|}, \tag{2}$$

where U represents the set of keywords associated with document A and V represents the set of keywords associated with document B at the intracluster level. At the intercluster level, U will represent the set of all keywords associated with the cluster C_i and V will represent the set of all keywords associated with the cluster C_j. This will be plotted against cluster size to

get a proxy measure of precision versus recall.

4.5 Project Timeline

The proposed research is expected to span a period of five years. Given the computational complexity and the iterative process of cluster calibration, we will dedicate approximately two years towards iteratively evaluating the initial quality of all networks at different resolutions. We will evaluate the viability of a given clustering for further analysis using the marker node method described below, as well as some internal cluster metrics. At the end of the 18th month, we will re-evaluate if any citation networks should be excluded from further work.

Months	**Activity**
1-18	Data Cleaning and Calculations
19 - 24	Initial Clustering and Cluster Calibrations
24 - 36	DC, BC, CC, DC-BC-CC-mLC Network Clustering
36 - 42	Quantitative Evaluation of Different Resolution Clustering
36 - 54	Qualitative Evaluation of Different Resolution Clustering
54 - 60	Further Analysis and Publication Preparation

Table 1: Expected Project Timeline: The data cleaning and edge weight calculation for a large dataset will be conducted over 18 months. At the end of this period, we anticipate selecting different networks based on their initial clustering quality. After calibration, the network clustering for different networks can be done concurrently. We will also be iteratively evaluating the partitions to refine our parameters. The quantitative and qualitative evaluation of the different clusterings will be also be conducted concurrently.

5 Future Directions

The results of this study will be foundational to achieving the long-term goal of developing citation cluster-based IR tools that serve the information needs of the scientific community. One next step would be determining how the journal information and the author information from Dimensions can be utilized to enhance the quality of the citation clusters. Then, the next phase will focus on creating an IR simulation tool as a cost-effective way of testing the performance of citation clusters without the added variable of user-behavior.

In the case that we discover our citation cluster-based hypothesis is infeasible or does not have sufficient merit to pursue further, we will still have contributed valuable information of citation cluster patterns on a significantly large dataset. Further, we will have evaluated the use of a new citation relationship and the patterns it may reveal within the structure of science. Scientometrics is still a burgeoning field and we believe our work will make meaningful additions to the body of work examining the structure and mapping of science.

6 Limitations

A certain level of risk is inherent in research involving community detection due to the lack of a ground-truth and the possibility that clustering algorithms may result in arbitrary communities. We mitigate some of this risk by utilizing both internal and external validators, as well as independent evaluation criteria in the form of keywords. We remain confident that our methods should provide some reliable insight into what parameters yield clusters of high similarity. However, in case we find that we are unable to maintain high node coverage of at least 80% or the quality of our clusters are poor, we will also evaluate the use of alternate clustering algorithms such as Iterative *k-Core* (IKC) and Infomap.

Another limitation inherent to working with a complex, dynamic network is accounting for the frequency with which it changes when creating whole corpus-based clusters. We believe that in a production setting, the computational cost of frequent clustering can be offset by the value the tool brings to the community. Further investigations are, of course, necessary.

7 Budget

		Amount
Personnel		
Principal Investigator Salary	Effort 100%	$50,000
Graduate Students	3 x $25/hr, 30hr/wk	$117,000
Database Manager	Effort 20%	$40,000
Software Engineer	Effort 30%	$40,000
Research Collaborator	Haadi Elsaawy, Effort 100%	$50,000
Statistician	Effort 5%	$20,000
Research and Support		
Server*		$30,000
Travel and Conference Expenses		$10,000
Citation Data		$30,000
Publication and Dissemination		$5,000
Total Budget (first year expenses)		**$392,000**
Total Budget (per annum, year 2 - 5)		**$362,000**
Total Budget (Five Years)		**$1,840,000**

Table 2: This is an estimated breakdown of the per annum expenses.
*Listed server expenses include yearly maintenance fee approximated to total $10,000.

References

[1] Per Ahlgren, Yunwei Chen, Cristian Colliander, and Nees Jan van Eck. Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications. *Quantitative Science Studies*, 1(2):714–729, 06 2020.

[2] Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck, and Ludo Waltman. Academic information retrieval using citation clusters: in-depth evaluation based on systematic reviews. *Scientometrics*, 128(5):2895–2921, 2023.

[3] Elmer V Bernstam, Jorge R Herskovic, Yindalon Aphinyanaphongs, Constantin F Aliferis, Madurai G Sriram, and William R Hersh. Using citation data to improve retrieval from medline. *J Am Med Inform Assoc*, 13(1):96–105, February 2006.

[4] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224, 2021.

[5] Kevin W. Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404, 2010.

[6] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5):512–546, 2011.

[7] Jochen Gläser, Wolfgang Glänzel, and Andrea Scharnhorst. Same data—different results? towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2):981–998, 2017.

[8] Michael Gusenbauer and Neal R. Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217, 2023/07/20 2020.

[9] Matthias Held, Grit Laudel, and Jochen Gläser. Challenges to the validity of topic reconstruction. *Scientometrics*, 126(5):4511–4536, 2021.

[10] Daniel W. Hook, Simon J. Porter, and Christian Herzog. Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 2018.

[11] Vidya Kamath, Vikram Ramavarapu, Fabio Ayres, and George Chacko. Connectivity modifier pipeline. `https://github.com/illinois-or-research-analytics/cm_pipeline`, 2023.

[12] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

[13] Richard Klavans and Kevin W. Boyack. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4):984–998, 2017.

[14] Robert Krovetz and W. Bruce Croft. Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10(2):115–141, April 1992.

[15] Anton Leuski. Evaluating document clustering for interactive information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 33–40, New York, NY, USA, 2001. Association for Computing Machinery.

[16] Minhyuk Park, Eleanor Wedell, Dima Korobskiy, Tandy Warnow, and George Chacko. Community finding and clustering project. Github repository, University of Illinois Urbana-Champaign, 2021.

[17] Taemin Kim Park. The nature of relevance in information retrieval: An empirical study. *The Library Quarterly: Information, Community, Policy*, 63(3):318–351, July 1993.

[18] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Santo Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734–745, October 2015.

[19] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[20] H. Small. Update on science mapping: Creating large document spaces. *Scientometrics*, 38(2):275–293, February 1997.

[21] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, August 1973.

[22] Henry Small. Navigating the Citation Network. *Proceedings of the ASIS Annual Meeting*, 32:118–26, 1995. ERIC Number: EJ513864.

[23] Henry Small. A Passage Through Science: Crossing Disciplinary Boundaries. *Library Trends*, 1999. Publisher: Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign.

[24] Lovro Šubelj, Nees Jan van Eck, and Ludo Waltman. Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE*, 11(4):e0154404–, 04 2016.

[25] Vincent Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

[26] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurence data in information retrieval. *Journal of Documentation*, 33(2):106–119, 1977.

[27] Ulrike Von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, pages 65–79. JMLR.org, 2011.

[28] Eleanor Wedell, Minhyuk Park, Dmitriy Korobskiy, Tandy Warnow, and George Chacko. Center–periphery structure in research communities. *Quantitative Science Studies*, 3(1):289–314, April 2022.

[29] Peter Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577–597, 1988.

[30] Dietmar Wolfram. The symbiotic relationship between information retrieval and informetrics. *Scientometrics*, 102(3):2201–2214, 2015.