Chapter 1

Introduction

Navigating academic literature has never been easy, and it is becoming harder each year due to the accelerating rate of literature production [6, 26, 115]. Traditional search methods, such as keyword-based queries, work well for users familiar with the topic, but leave them unaware of blind spots [106, 169]. On the other hand, science maps are tools to visualize the relationships between academic documents, which function as an overview of the research landscape. Science maps can support traditional search by showing overlooked connections and allowing expert knowledge to improve the search. Despite this, science maps are not part of information seeking manuals or studies [32, 95]. Given this situation, it would be beneficial to get a better understanding of the performance of science maps for information retrieval tasks. In this dissertation, we do this by providing evidence for the benefits and drawbacks of using science maps for specific information retrieval tasks, and by exploring ways for using them more effectively.

1.1 Use of science maps

Science maps are tools to visually explore the relations between objects of interest in a large collection of documents [25, 38, 127]. Their most typical use is in studying the structure of a scientific field. For example, each year there are countless bibliometrics studies of different academic fields that have science maps as their core method of study. Among other things, these studies delimit the field [185], identify topical trends, research groups, and key actors and events. The target audience of these studies are typically members of the same field, as it allows them to know where their research stands in relation to their colleagues. Beyond studying the structure of a scientific field, policymakers and research administrators also use science maps to characterize national research output, evaluate the impact of funding programs, and identify competitive advantages and weaknesses in different scientific areas.

Science maps belong to a broader family of tools that attempt to represent and summarize large collections of data in a structured and interpretable manner. These tools are used both for analysis and communication. For example, Latent Dirichlet Allocation [24], a method to algorithmically generate topic models, is used to identify topics in documents according to the co-occurence of words in documents. Similarly, conceptual maps [58], which are manually constructed diagrams of relations between concepts, are used in education to explain how concepts relate to each other.

Science map visualizations are typically a bubble chart that represents a network, where the bubbles (i.e. nodes) represent objects of interest. These nodes are typically either academic researchers, journals, organizations, countries, terms from the documents, or clusters of documents with a topical label. Figure 1.1 is an example of a science map that visualizes clusters of documents, and Figure 1.2 is an example of a science map that visualizes authors.

The bubbles (nodes) in a science map are placed in a two dimensional plane, sometimes connected by the network edges, where the spatial position and edges represent the relation of the nodes with

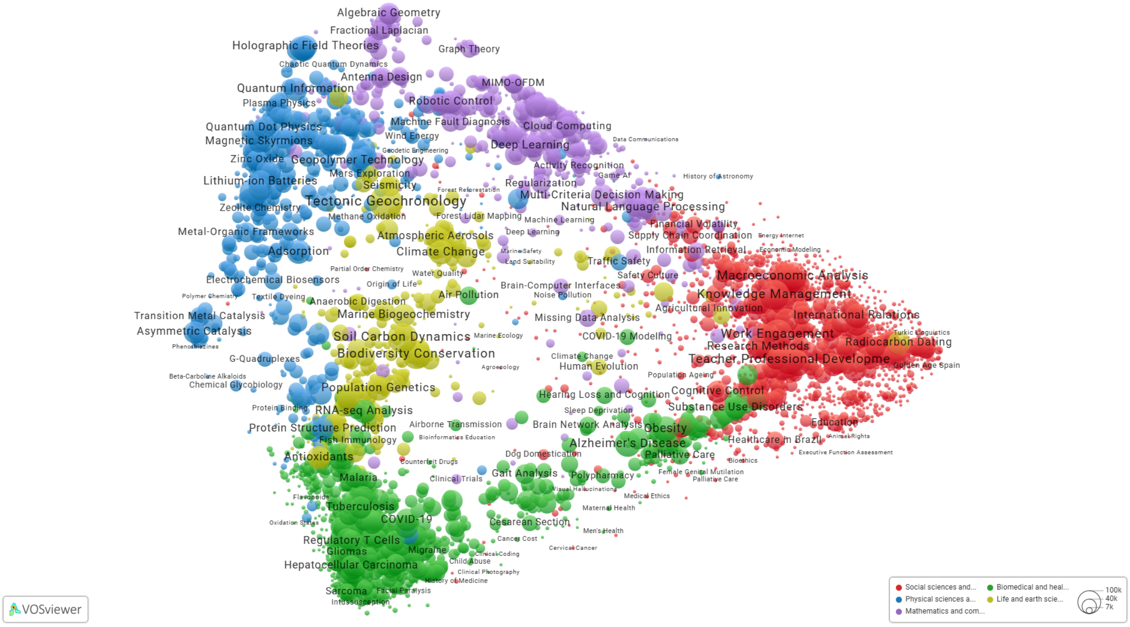


Figure 1.1: Example of a science map that visualizes clusters of documents. Spatial proximity indicates the intensity of intercluster citations, and color indicates the field of science. Source [159]

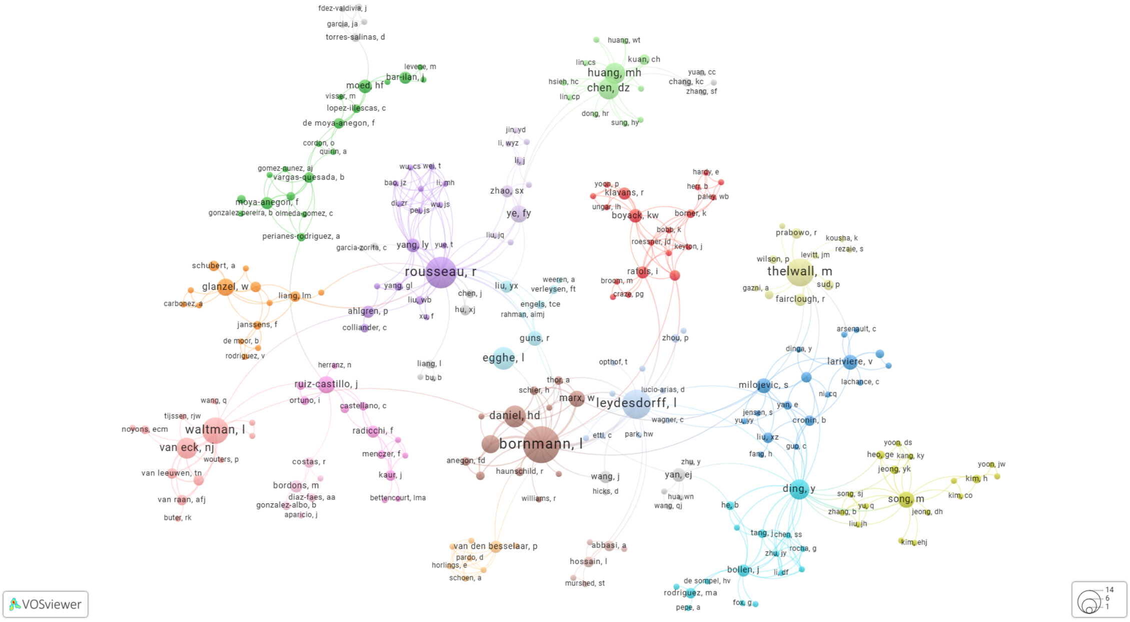


Figure 1.2: Example of a science map that visualizes authors. Links and spatial proximity indicate co-authorships, and color indicates that they belong to the same cluster in the network. Source [155]

each other. Other properties, such as the color and size, are used to represent other properties of the node. This representation allows users of science maps to quickly find connections between the nodes. For example, authors that are placed together (Figure 1.2) probably belong to the same research group or work on the same topic.

1.2 Document clusters in science maps

There are many types of science maps, but in this dissertation we only focus on the ones that use clusters of documents as nodes in their visualizations [38, 151, 164]. We do this because the use of these maps in information retrieval is straight-foward, as a user can just retrieve the documents that belong to a given cluster. Typically, the clusters are created the following way: First, the collection of documents to be visualized by the science map is turned into a network, where the nodes are the documents and the links are based on the metadata of the documents (see below). Then, a clustering algorithm is run to find communities of documents in the network. A frequently used algorithm is the Leiden algorithm, which requires a resolution value to be provided to determine the granularity of the clusters [153]. These two steps are the ones we study the most in this dissertation. Then, the next step before visualization is to assign labels to the clusters, which typically is done with conventional text processing techniques (like word count vectors [111]) rather than advanced (like text embeddings [50]) or field-specific ones (like named entity recognition of diseases in health records [168]). For example, Waltman and van Eck [164] did the former, while van Eck and Waltman [159] did the latter. Finally, the clusters are visualized as a network of clusters using the approach described in Section 1.1.

There are three types of networks that are typically used for document clustering. The first and most common type is the citation network. In its most simple implementation, called direct citations network, the edges of the network are the citation links between documents. Using citations has the advantage that a citation is an explicit intellectual link made by academics. There are three other implementations of citation networks:

- Extended direct citation network: It includes citations to documents that are not part of the science map [165].
- Co-citation network: An edge in the network indicates that the two linked documents are cited by the same third document [145].
- Bibliographic coupling network: An edge in the network indicates that the two linked documents cite the same third document [96].

The edges of co-citation and bibliographic coupling networks can have different weights according to how many documents they are being cited by or are citing together.

The second type of network used for document clustering is the similarity network, where similarity typically is text similarity. The weight of the edge between two documents is the strength of the text similarity, and the text used to determine the text similarity typically consists of the titles and abstracts of the documents because this data tends to be readily available. The methods used to calculate the similarity are diverse, but it is notable that advanced methods, like text embedding, tend to not be used, likely due to bibliometricians preferring explainability over performance. The text similarity network has an advantage over the citation network in that the similarity between any pair of documents can be calculated, which helps when citation links between the documents are sparse, for example citation networks where the documents are very new or very few. There are also hybrid networks where both citations and text similarity contribute to the weights of the edges [4, 27].

The third type of network is the co-occurrence network. In this network, an edge between two documents indicates that the documents share something, like an author, or are placed together in some context, like being cited by a patent. It is worth mentioning that citation and text networks can also be seen as co-occurrence networks (like in the previously mentioned co-citation and bibliographic coupling, or by sharing a word in the text). Beyond these, some of the most common elements used to connect documents in a co-occurrence network are institutions, policy documents, social media posts and key words [45, 101]. Co-occurrence networks are used to answer specific questions related to the entities that connect the documents, for instance how social media connects academic documents.

1.3 Information retrieval with clusters

The field of information retrieval studies how to find relevant information. Over time, research in this field has shifted away from using clustering to find information due to the emergence of well performing ranking algorithms. The pivotal point came with the emergence of PageRank [29], from Google, which was able to rank web pages not only according to the relevance of their textual content to a query, but also based on the number and weight of the hyperlinks pointing at it. Even so, cluster-based information retrieval should not be ignored because there are tasks that are better served by clustering than by a ranked list of results [176]. One of the most relevant uses is diversification of search results [72]: In the example of PageRank, it is possible for a query to match several topics. If only one of the topics appears at the top of the result list, then there is the risk that this is not the topic that the user is looking for. For example, if the user is looking for the homepage of the company Jaguar, but all the results are about the animal. With results diversification, the results can be clustered by topic and then ensure that all the topics are represented at the top of the list. The search engine Carrot2 goes one step further [65] and makes the results topic clusters available to the user. Another use of clusters is query expansion [108], where the user does not know all the relevant query terms for the search and then the system suggests new terms based on the ones that the user already provided.

As a general rule, clustering in information retrieval is most useful at assisting in complex tasks that require several steps, instead of simple tasks like finding a known individual document [71]. These complex tasks require the user to explore the results of each step so as to decide on the next step. Clustering also supports complex tasks by giving tools to the users to expand or reduce their search results in a sensible way. An excellent example of clustering for complex multi-step tasks is the Scatter-Gather method [48], which was originally proposed to facilitate the navigation of news articles. It creates clusters based on the text similarity between the documents, then labels the clusters, and then lets the user select which clusters might contain documents with their topic of interest. Then the method creates a new set of clusters using only the documents in the previously selected clusters, and then the process repeats until the user identifies a cluster labeled with their topic of interest. This approach is ideal when the user struggles to articulate effectively their topic of interest, when they do not know how to find their topic of interest in the documents, or when the documents categorization is lacking.

In this dissertation, we hypothesize that the Scatter-Gather method is ideal for academic search for two reasons:

- Because academic users tend to have complex information needs [136].
- Because academic documents categorization tends to be lacking due to the rate of emergence of new topics and research questions, and the difficulty to maintain a classification system [3, 180].

1.4 Bibliometrics enhanced information retrieval

Academics have used bibliometrics to enhance information retrieval through the use of citations [30, 63, 114]. The closest to the Scatter-Gather method is the tool CitNetExplorer [157]. This tool creates a network of citations between documents, and it can also identify clusters within the network. Its user interface facilitates selection of documents based on the clusters they belong to. It can also create a new clustering solution based on selected documents, including documents selected using the clusters, which allows the users to easily follow the Scatter-Gather method. It also facilitates selecting additional documents based on their citation links to the currently selected documents, something that is arguably a helpful addition because the Scatter-Gather method only allows to remove documents. However, instead of cluster-based methods, academics tend to use citations for information retrieval using a method called citation snowballing [23]. Snowballing consists of selecting one or more documents, gathering the documents that cite them or are cited by them, and then repeating the process with this new set of documents until the user is satisfied. To limit the number of selected documents, there is usually a limit on how many expansion cycles to make or a minimum threshold on the number of selected documents a new document must be connected to. For example, Janssens and Gwinn [89] found that, for finding the relevant documents of systematic reviews, it was most convenient to add documents that are co-cited by multiple seed documents and to use the number of co-citations as a threshold for relevance.

Additionally, there are tools that visualize the structure of search results, either by text similarity (like Open knowledge maps [122] and Iris.ai [88]) or citation (like Inciteful [87], Litmaps [107], Connected papers [44] and Research rabbit [99]). A difference between these tools and science maps is that these tools operate at a very small scale, visualizing individual documents. This has the advantage that the tools work in a way that is intuitive and easy-to-understand for users, but it misses the big picture view that is provided by working with big clusters of documents. For example, most of the clusters in Figure 1.1 have between 1,000 and 100,000 documents. CitNetExplorer found a middle point between visualizing clusters and visualizing individual documents by visualizing only the most important documents, which are identified by the properties of their nodes in the citation network. The clusters are kept in the back end of the system and the visualization indicates the clusters of the visualized documents. Another difference between science maps and the above-mentioned tools is that the latter tend to create their network of documents starting from seed documents provided by the users, which does not allow the users to know what they are missing. This is particularly problematic where there are communities of relevant documents disconnected from each other, either by lack of citations or lack of similar language [2, 80, 134]. For example, the collection of documents about academic information retrieval has a low citation connectivity between its biomedicine and computer science communities, and a user might never be aware of this if they use the above-mentioned tools with seed documents from one community only.

In summary, information retrieval based on science maps uses similar concepts and methods to bibliometric enhanced information retrieval, but it also has potential advantages for information retrieval that differentiates it from the latter. Given this, we identify the lack of knowledge on the performance of science maps for information retrieval as a research gap that we will attempt to fill in this dissertation.

1.5 Research questions

Our motivation for the research presented in this dissertation is to explore the potential of science maps to enhance academic search by saving time, improving knowledge discovery, and providing a more complete picture of the state of the art in the literature. The benefits of science maps may be especially significant for early career researchers and citizen scientists, who tend to be less familiar with their research field. Our research also adds value to the field of science mapping because it provides a new application area for the knowledge generated by the field. The way we see it, a research agenda for this purpose should evaluate the effectiveness of science maps for information retrieval tasks and also improve this effectiveness. Therefore, in this dissertation, our overarching research question is: **What is the effectiveness of science maps for information retrieval, and how can we enhance it?** We address this overarching research question by answering a number of more specific research questions, and each of them is the topic of a separate chapter in this thesis:

- Research question 1 (Chapter 2): The first step in our research is to understand how to use science maps for information retrieval. Our research question is: **How can science maps be designed to support information retrieval?** Here, we propose a system, named SciMacro

(Science Macroscope), for interacting with science maps that serves information retrieval tasks based on the Scatter-Gather method. We find no significant hindrances for the implementation. Figure 1.3 shows a screenshot of the graphical user interface that we created for SciMacro, whose source code is publicly available [15].

- Research question 2 (Chapter 3): To evaluate science maps for information retrieval tasks, we start by addressing the research question: **How effective are science maps for making systematic reviews?** Here, we evaluate the performance of science maps at retrieving the relevant documents of systematic reviews, using the Boolean queries of systematic reviews as baseline. We find that science maps are able to outperform the baseline for about half of the systematic reviews.
- Research question 3 (Chapter 4): We also consider whether the performance of a science map depends on the academic topic of the task. Our research question is: **Do science maps represent some topics better than others?** Here, we evaluate the performance of science maps at creating clusters for topics, using Medical Subject Headings as topics. We find that both text and citation based maps cluster the topics that belong to ontological categories of topics "Organisms" or "Diseases" much better than the other topics.
- Research question 4 (Chapter 5): Finally, we consider whether we can manipulate a science map to perform better at a given academic topic. Our research question is: **How can the representation of specific topics be improved in a science map?** Here, we evaluate if the use of different kinds of document networks influences which topics are clustered better than others. We find that such an influence indeed exists, but also that the topics from most ontological categories of topics decreased their performance in the new networks relative to text or citation networks and that performance can be improved by merging different network types.

Research questions 1 and 4 investigate the use and improvement of science maps for information retrieval, while research questions 2 and 3 evaluate their effectiveness. The progression is as follows: First, we conceptualize how science maps can support information retrieval (RQ1). Based on this, we then evaluate their performance (RQ2) and find it to be uneven. To understand this variation, we examine whether topic differences play a role (RQ3) and confirm that they do. Finally, we investigate whether performance for underrepresented topics can be improved by changing the data source of the map (RQ4). We conclude this dissertation in Chapter 6, where we summarize key findings and discuss future research directions for science maps for information retrieval. It is worth noting that this dissertation does not include experiments with real users. Although such experiments could help answer the research questions, they would require resources and expertise beyond the scope of this dissertation.

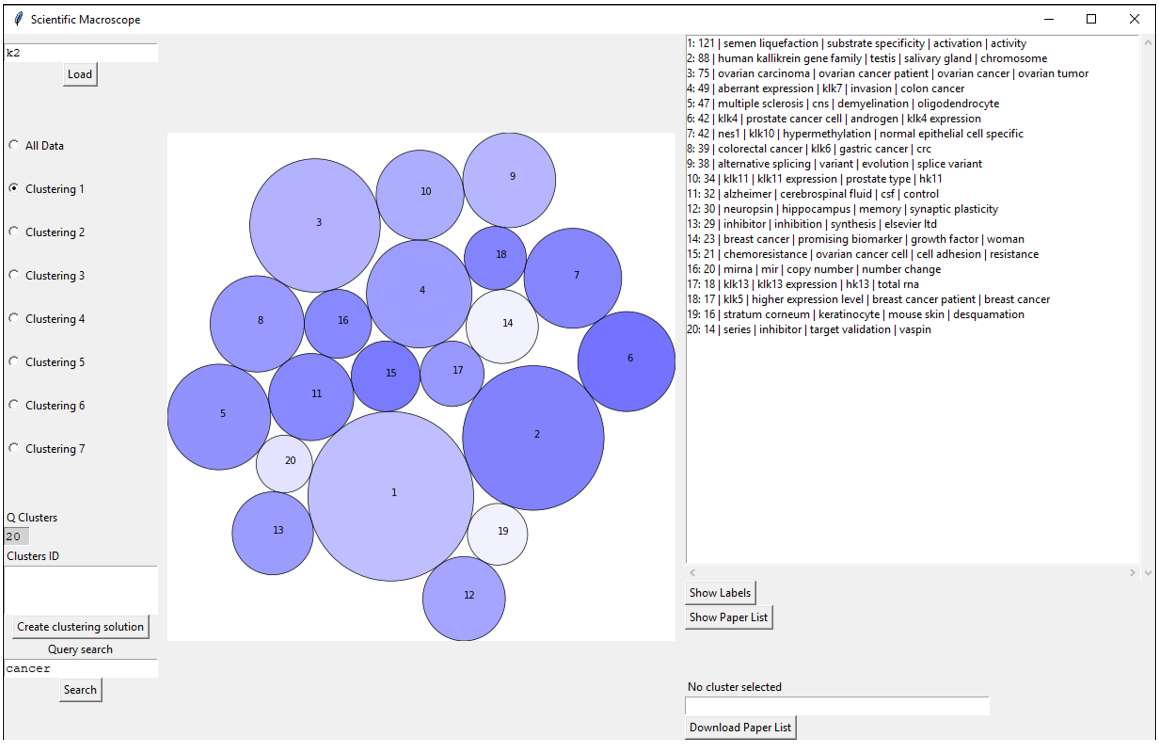
1.6 Main contributions

While this dissertation focuses on answering the research questions, we also made other additional contributions to the field. We divide the contributions between resource contributions and methods contributions.

1.6.1 Resource contributions

These are the resources that we generated during our dissertation, and they facilitate either the design of science mapping tools or the execution of information retrieval experiments:

- In Chapter 2, we introduce a science mapping tool prototype that follows the Scatter-Gather principles, including an algorithm that places the bubbles close to each other while also preventing overlapping and minimizing empty space. We have made the code publicly available [15].



Scientific Macroscope

k2

○ All Data

● Clustering 1

○ Clustering 2

○ Clustering 3

○ Clustering 4

○ Clustering 5

○ Clustering 6

○ Clustering 7

Q Clusters: 20

Clusters ID:

Query search: cancer

```
1: 121 | semen liquefaction | substrate specificity | activation | activity
2: 88 | human kallikrein gene family | testis | salivary gland | chromosome
3: 75 | ovarian carcinoma | ovarian cancer patient | ovarian cancer | ovarian tumor
4: 49 | aberrant expression | klk7 | invasion | colon cancer
5: 47 | multiple sclerosis | cns | demyelination | oligodendrocyte
6: 42 | klk4 | prostate cancer cell | androgen | klk4 expression
7: 42 | nes1 | klk10 | hypermethylation | normal epithelial cell specific
8: 39 | colorectal cancer | klk6 | gastric cancer | crc
9: 38 | alternative splicing | variant | evolution | splice variant
10: 34 | klk11 | klk11 expression | prostate type | hk11
11: 32 | alzheimer | cerebrospinal fluid | csf | control
12: 30 | neuropsin | hippocampus | memory | synaptic plasticity
13: 29 | inhibitor | inhibition | synthesis | elsevier ltd
14: 23 | breast cancer | promising biomarker | growth factor | woman
15: 21 | chemoresistance | ovarian cancer cell | cell adhesion | resistance
16: 20 | mirna | mir | copy number | number change
17: 18 | klk13 | klk13 expression | hk13 | total rna
18: 17 | klk5 | higher expression level | breast cancer patient | breast cancer
19: 16 | stratum corneum | keratinocyte | mouse skin | desquamation
20: 14 | series | inhibitor | target validation | vaspin
```

No cluster selected

Figure 1.3: Screenshot of the graphical user interface of SciMacro. On the left it allows the user to load the documents they want to cluster, indicate how many clusters they want, select clusters for the next clustering step, go back and forth between the clustering steps and search for words in the clusters whose frequency is represented by the color of the clusters. On the right it provides a description of the clusters, shows the documents inside the clusters, and allows downloading the documents in the selected cluster. Source code of the interface [15]

- In Chapter 3, we have developed [14] a manually curated version of the search queries dataset from Scells et al. [139] that is compatible with the PubMed API query grammar and that can be used for information retrieval experiments.
- In Chapters 3, 4 and 5, the source code of all the experiments is provided, along with a modified version of the experimental data that allows replication of the experiments. The data is modified to prevent legal violations.

1.6.2 Methodological contributions

In the current dissertation we evaluated the performance of science maps for information retrieval and explored ways of improving it.

- We proposed a user model that assumes a user that has perfect knowledge about the location of relevant documents, and use it in Chapter 3, 4 and 5 to design original evaluation methods.
- We proposed two approaches to select which are the relevant documents that should be used to evaluate science maps for different tasks: In Chapter 3, we proposed to use the included and excluded documents in a systematic review. In Chapter 4 and 5, we proposed to use the documents labeled with MeSH terms to evaluate academic topics and allowed the documents to inherit MeSH terms higher up in the ontology.
- We proposed how to evaluate the clusters in science maps in a way that manages the size disparity between clusters and number of relevant documents. In Chapter 3, we proposed to allow the user model to manage the granularity of the clusters. In Chapter 4 and 5, we proposed a method that aggregates the evaluations of multiple granularities.

1.7 List of publications

Chapters 2 to 5 in this thesis are based on the following publications.

Chapter 2: Juan Pablo Bascur, Nees Jan van Eck and Ludo Waltman. 2019. An interactive visual tool for scientific literature search: Proposal and algorithmic specification. Proceedings of the 8th International Workshop on Bibliometric-Enhanced Information Retrieval (BIR) Co-Located with the 41st European Conference on Information Retrieval (ECIR 2019), 76–87. https://ceur-ws.org/Vol-2345/paper7.pdf [16]

CRediT author statement:

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Chapter 3: Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2023. Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. Scientometrics, 128, 2895–2921. https://doi.org/10.1007/s11192-023-04681-x [17]

CRediT author statement:

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Software, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Chapter 4: Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2025. Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks. Scientometrics 130, 1181–1199. https://doi.org/10.1007/s11192-024-05218-6 [19]

CRediT author statement:

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Chapter 5: Juan Pablo Bascur, Rodrigo Costas, Suzan Verberne. 2024. Use of diverse data sources to control which topics emerge in a science map. arXiv. https://doi.org/10.48550/arXiv.2412.07550 [18]

CRediT author statement:

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Rodrigo Costas: Conceptualization, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.