



Universiteit  
Leiden  
The Netherlands

## Science maps for information retrieval

Bascur Cifuentes, J.P.

### Citation

Bascur Cifuentes, J. P. (2026, January 21). *Science maps for information retrieval*. Retrieved from <https://hdl.handle.net/1887/4287774>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4287774>

**Note:** To cite this publication please use the final published version (if applicable).

---

# Science maps for information retrieval

---

by

Juan Pablo Bascur Cifuentes



Universiteit  
Leiden  
The Netherlands

Leiden, 2026

---



# Science maps for information retrieval

Proefschrift

ter verkrijging van  
de graad van doctor aan de Universiteit Leiden,  
op gezag van rector magnificus prof.dr. S. de Rijcke,  
volgens besluit van het college voor promoties  
te verdedigen op woensdag 21 januari 2026  
klokke 16:00 uur  
door  
Juan Pablo Bascur Cifuentes  
geboren te Providencia, Chile  
in 1987

**Promotores**

Prof.dr. L.R. Waltman

Prof.dr. S. Verberne

**Co-promotor:**

Dr. N.J.P. van Eck

**Promotiecommissie:**

Prof.dr. B.A. Barendregt (Voorzitter/Decaan Graduate School)

Prof.dr. R.J.W. Tijssen

Prof.dr. C.K. Kreutz (Technische Hochschule Mittelhessen)

Prof.dr. G. Cabanac (Université de Toulouse)

Dr. T. Velden (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung)

# Contents

<b>Acknowledgments</b>	<b>1</b>
<b>Quote</b>	<b>2</b>
<b>Summary in English</b>	<b>3</b>
<b>Summary in Dutch</b>	<b>5</b>
<b>About the author</b>	<b>7</b>
<b>1 Introduction</b>	<b>8</b>
1.1 Use of science maps . . . . .	8
1.2 Document clusters in science maps . . . . .	10
1.3 Information retrieval with clusters . . . . .	11
1.4 Bibliometrics enhanced information retrieval . . . . .	11
1.5 Research questions . . . . .	12
1.6 Main contributions . . . . .	13
1.6.1 Resource contributions . . . . .	13
1.6.2 Methodological contributions . . . . .	15
1.7 List of publications . . . . .	15
<b>2 An interactive visual tool for scientific literature search: Proposal and algorithmic specification</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.2 Description of the tool . . . . .	18
2.3 Case study of the tool . . . . .	19
2.3.1 Set up . . . . .	19
2.3.2 Example of the search process . . . . .	19
2.4 Technical specification . . . . .	20
2.4.1 Clustering the documents . . . . .	20
2.4.2 Labeling the clusters . . . . .	20
2.4.3 Visualizing the clusters . . . . .	23
2.5 Conclusion . . . . .	25
2.6 Data availability . . . . .	25
2.7 CRediT author statement . . . . .	25
2.8 Appendix . . . . .	25
<b>3 Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews</b>	<b>28</b>
3.1 Introduction . . . . .	28
3.2 Related work . . . . .	29
3.2.1 Science mapping . . . . .	29

3.2.2	Citation-based IR . . . . .	30
3.2.3	Cluster-based IR . . . . .	30
3.3	Method . . . . .	31
3.3.1	Task design and data collection . . . . .	31
3.3.2	Citation network . . . . .	32
3.3.3	Simulation of CCIR . . . . .	33
3.3.4	Quantitative analysis . . . . .	35
3.3.5	Qualitative analysis . . . . .	36
3.4	Results . . . . .	37
3.4.1	Quantitative results . . . . .	37
3.4.2	Qualitative results . . . . .	39
3.5	Discussion . . . . .	47
3.5.1	What types of users are best served by CCIR? . . . . .	47
3.5.2	What types of SRs are best served by CCIR? . . . . .	47
3.5.3	What are the strengths and weaknesses of CCIR? . . . . .	47
3.5.4	Limitations of this work . . . . .	48
3.6	Conclusion . . . . .	49
3.7	Data availability . . . . .	50
3.8	CRedit author statement . . . . .	50
3.9	Acknowledgements . . . . .	50
<b>4</b>	<b>Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks</b>	<b>51</b>
4.1	Introduction . . . . .	51
4.2	Background . . . . .	52
4.2.1	Evaluation of science maps . . . . .	52
4.2.2	Criticism of science maps based on ground truth evaluations . . . . .	53
4.2.3	Meaning of the clusters . . . . .	53
4.3	Methods . . . . .	54
4.3.1	Data selection . . . . .	54
4.3.2	Data preprocessing . . . . .	54
4.3.3	Clustering effectiveness . . . . .	57
4.4	Results . . . . .	58
4.4.1	Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks? . . . . .	58
4.4.2	Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa? . . . . .	58
4.5	Discussion . . . . .	59
4.5.1	Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks? . . . . .	60
4.5.2	Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa? . . . . .	61
4.5.3	Strengths and weaknesses . . . . .	62
4.6	Conclusion . . . . .	63
4.7	Data availability . . . . .	64
4.8	CRedit author statement . . . . .	64
<b>5</b>	<b>Use of diverse data sources to control which topics emerge in a science map</b>	<b>65</b>
5.1	Introduction . . . . .	65
5.2	Background . . . . .	67
5.2.1	Interaction of academic documents with non-academic elements . . . . .	67
5.2.2	Science maps based on diverse sources . . . . .	67
5.2.3	Criticisms to maps of science . . . . .	67

5.2.4	Comparing clustering solutions of different networks . . . . .	68
5.3	Methods . . . . .	68
5.3.1	Core academic documents . . . . .	68
5.3.2	External sources networks . . . . .	68
5.3.3	Text similarity networks . . . . .	70
5.3.4	Citation network . . . . .	70
5.3.5	Clustering . . . . .	71
5.3.6	Topics and topic categories . . . . .	71
5.3.7	Evaluation . . . . .	73
5.3.8	Summary of methods . . . . .	76
5.4	Results . . . . .	77
5.4.1	Citations . . . . .	82
5.4.2	Twitter conversations . . . . .	82
5.4.3	Document authors . . . . .	83
5.4.4	Facebook users . . . . .	83
5.4.5	Policy documents . . . . .	84
5.4.6	Patent families . . . . .	84
5.4.7	Twitter authors . . . . .	84
5.4.8	Twitter networks versus the other networks . . . . .	85
5.4.9	Cases where Purity decreases at higher NSC . . . . .	85
5.5	Discussion . . . . .	85
5.6	Conclusions . . . . .	87
5.7	Data availability . . . . .	88
5.8	CRedit author statement . . . . .	88
<b>6</b>	<b>Conclusion</b> . . . . .	<b>89</b>
6.1	Answers to research questions . . . . .	89
6.2	Further research . . . . .	90
	<b>Bibliography</b> . . . . .	<b>93</b>



# List of Figures

1.1	Example of a science map that visualizes clusters of documents. . . . .	9
1.2	Example of a science map that visualizes authors. . . . .	9
1.3	Screenshot of the graphical user interface of SciMacro. . . . .	14
2.1	The Scatter/Gather approach. . . . .	18
2.2	Visualization of clusters. . . . .	20
2.3	Illustration of the minimization algorithm. . . . .	24
3.1	Cluster selection algorithm. . . . .	35
3.2	Precision, Recall and F-Score. . . . .	38
3.3	Intersection proportions. . . . .	39
3.4	Documents sets sizes. . . . .	40
3.5	Tree-level of the retrieved clusters. . . . .	41
3.6	Venn diagram of the intersections. . . . .	42
4.1	Box plots showing the distribution of C-Purity, C-ICC, T-Purity and T-ICC over the 45 combinations of parameter values. . . . .	60
4.2	Box plots showing the distribution of rPurity and rICC for each value of Size bin, Resolution and Coverage. . . . .	61
4.3	Box plots showing the distribution of rPurity and rICC for each branch. . . . .	62
5.1	Example of a Purity profile. . . . .	74
5.2	Diagram on the representation of results. . . . .	75
5.3	Examples of Purity of several topic categories for different networks. . . . .	80
5.4	Examples of Purity profiles for individual topics across different networks. . . . .	81

# List of Tables

2.1	Labels of the first scattering. . . . .	21
2.2	Labels of the second scattering. . . . .	21
2.3	Top 5 papers for cluster 1 in the second scattering. . . . .	22
2.4	Top 5 papers for cluster 2 in the second scattering. . . . .	22
2.5	Top 5 papers for cluster 4 in the second scattering. . . . .	23
3.1	Quantitative data of the SRs in the qualitative analysis. . . . .	43
3.2	Characterization of the SRs. . . . .	44
3.3	Topic of the sets of documents of the SRs. . . . .	45
4.1	Statistics of the clustering solutions. . . . .	55
4.2	Number of MeSH terms per branch and Size bin. . . . .	56
4.3	Number of times each branch appears in each ranking position, using either C-Purity (top) or T-Purity (bottom) as ranking criterion. . . . .	59
5.1	List of topic categories used in the current paper. . . . .	72
5.2	Size bins per source after filtering. . . . .	72
5.3	Detail of the results of each network. . . . .	78
5.4	Summary of the results for each network. . . . .	79
5.5	Best (non-citation) networks per topic category from Table 5.4. . . . .	82

# Acknowledgments

Doing a PhD has been a life-changing experience. During these eight years, I have struggled, grown, and changed in ways I could never have imagined. But I could not have done this alone. Throughout this time, I have been supported, trusted, and taught by people who helped me learn things I could never have learned by myself.

I have so much to be grateful for, so I decided to organize this in a more manageable and structured way:

- **Ludo Waltman** and **Nees Jan van Eck**, who accepted me as a PhD student and had faith in my abilities.
- **Suzan Verberne**, who joined as my supervisor, believed in me, and provided critical support throughout my research.
- **Leiden University**, for making my burnout recovery easier.
- My partners **Roel van der Ploeg** and **Pepijn Stoop**, for their constant support during this journey.
- The **State of the Netherlands**, for giving me the opportunity to emigrate and integrate into a new society where I feel much more comfortable with the culture, the people, and the systems, and where I have found far greater opportunities for professional growth.
- **Suzan Verberne**, **Rodrigo Costas**, **Vincent Traag**, and **Ed Noijons**, for helping me find paid positions, which I desperately needed to sustain my self-funded PhD.
- For their rich academic discussions that shaped my research, I thank **Ludo Waltman**, **Nees Jan van Eck**, **Suzan Verberne**, **Rodrigo Costas**, **Vincent Traag**, **Ismael Rafols**, **Alfredo Yegros**, **Theo van Leeuwen**, **Martijn Visser**, **Jonathan Dudek**, **Wout Lamers**, and **Qianqian Xie**.
- **Karin den Dulk** and **Petra van der Weel**, for their kind support with administrative matters.
- Finding a postdoctoral position is never easy, and I am grateful to **Maxime Holmberg Sainte-Marie** for his support during that search.

During these years, I had the privilege of meeting many people in the academic environment whose friendship and companionship gave meaning and joy to my work. I am grateful to:

- My fellow CWTS researchers **Sonia Mena**, **Jonathan Dudek**, and **Mark Neijssel**.
- My fellow CWTS and LIACS PhD colleagues **Zhichao Fang**, **Ana Parrón**, **Wout Lamers**, **Alex Brandsen**, **André Brasil**, **Arian Askari**, **David Lindevelt**, and especially **Qianqian Xie**.
- My fellow visiting scholars **Zhentaoyang Liang**, **Fabiano Borges**, **Gianmarco Spinaci**, **Thamyres Choji**, **Honami Numajiri**, **Jingwen Zhang**, and especially **Alause Pires**.
- My colleagues from the wider academic community **Wenjing Chu**, **Jing Wang**, **Daniëlle van Rijk**, **Felipe Castillo**, **Julia Eckert**, **Hui Ching Hung**, **Eveline de Boer**, **Qinyu Chen**, **Satya Almasian**, and **Ylenia Casali**.

There is a world of people I am not mentioning simply because I cannot recall everyone right now. But please know that even if I have not named you, your actions have had a huge impact on my life. **Thank you.**

# Quote

*We are drowning in information but starved for knowledge*  
— John Naisbitt, Megatrends