Preliminary Results of a Scientometric Analysis of the German Information Retrieval Community 2020-2023

Philipp Schaer, Svetlana Myshkina, and Jüri Keller
Technische Hochschule Köln, Cologne, Germany

Version: 2023-10-10

Outline

German Information Retrieval community

- Information Science AND Computer Science
- there are no current studies that investigate these communities on a scientometric level.
- available studies only focus on the information scientific part of the community.

Data Set

- 401 recent (2020-2023) IR-related publications extracted from six core IR conferences from a mainly computer scientific background.

Analysis

- institutional level and
- researcher level

Use Cases

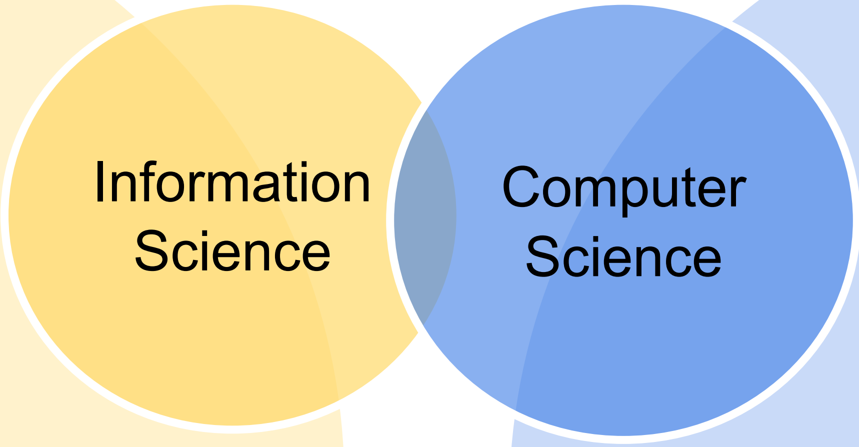
IR is heterogeneous

Hochschulverband Informationswissenschaft
100 members

Conferences
- ISI
- iConference
- ISIC
- ASIS&T annual meeting

…

Baumgartner and Schlögl [2]
Lewandowski and Haustein [3]
Larivière et al. [4]



Information Science

Computer Science

Fachgruppe IR (Gesellschaft für Informatik)
200 members

Conferences
SIGIR
ECIR
CHIIR
CIKM
CLEF
WWW
…

Ding et al. [1]
Thornley et al. [5]
Larsen [6]

Bibliometric insights into IR

Baumgartner and Schlögl [2] → only IS

- 1990 and 2004, ISI conference proceedings
- 1.6 authors per paper / 81% German papers / Konstanz, Graz, Regensburg, Hildesheim, and Saarbrücken

Lewandowski and Haustein [3] → only IS

- Handbook “Grundlagen der praktischen Information und Dokumentation”
- 1.4 authors per paper

Ding et al. [1]: analysis from 2000 → too old

Thornley et al. [5] → only TRECVid

Larsen [6]: → only CLEF

No recent dataset / paper on German IR community

Data Set

Six major **IR-related** and **peer-reviewed** conferences

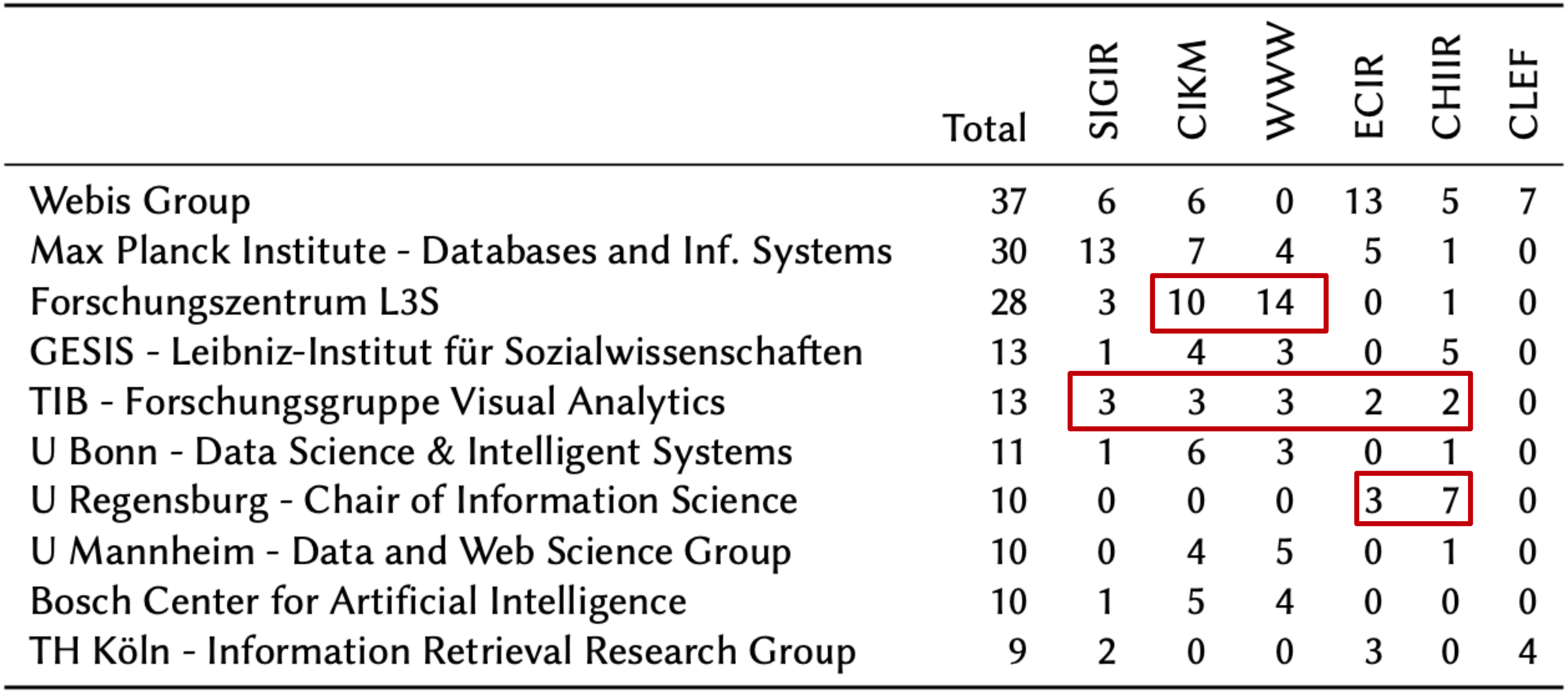
- CHIIR, CIKM, CLEF, ECIR, SIGIR, WWW.
- Published between January 2020 and June 2023
- At least one German author or an author that was affiliated with a German research institute
- Not included: Journals, like Information Retrieval Journal or ACM TOIS

Metadata for 401 publications:

- author names,
- affiliations (195 distinct),
- titles,
- DOI of the publication.

https://github.com/irgroup/LWDA2023-IR-community

Most productive IR research groups



	Total	SIGIR	CIKM	WWW	ECIR	CHIIR	CLEF
Webis Group	37	6	6	0	13	5	7
Max Planck Institute - Databases and Inf. Systems	30	13	7	4	5	1	0
Forschungszentrum L3S	28	3	10	14	0	1	0
GESIS - Leibniz-Institut für Sozialwissenschaften	13	1	4	3	0	5	0
TIB - Forschungsgruppe Visual Analytics	13	3	3	3	2	2	0
U Bonn - Data Science & Intelligent Systems	11	1	6	3	0	1	0
U Regensburg - Chair of Information Science	10	0	0	0	3	7	0
U Mannheim - Data and Web Science Group	10	0	4	5	0	1	0
Bosch Center for Artificial Intelligence	10	1	5	4	0	0	0
TH Köln - Information Retrieval Research Group	9	2	0	0	3	0	4

Most productive IR research groups

Mixed set of publication profiles.

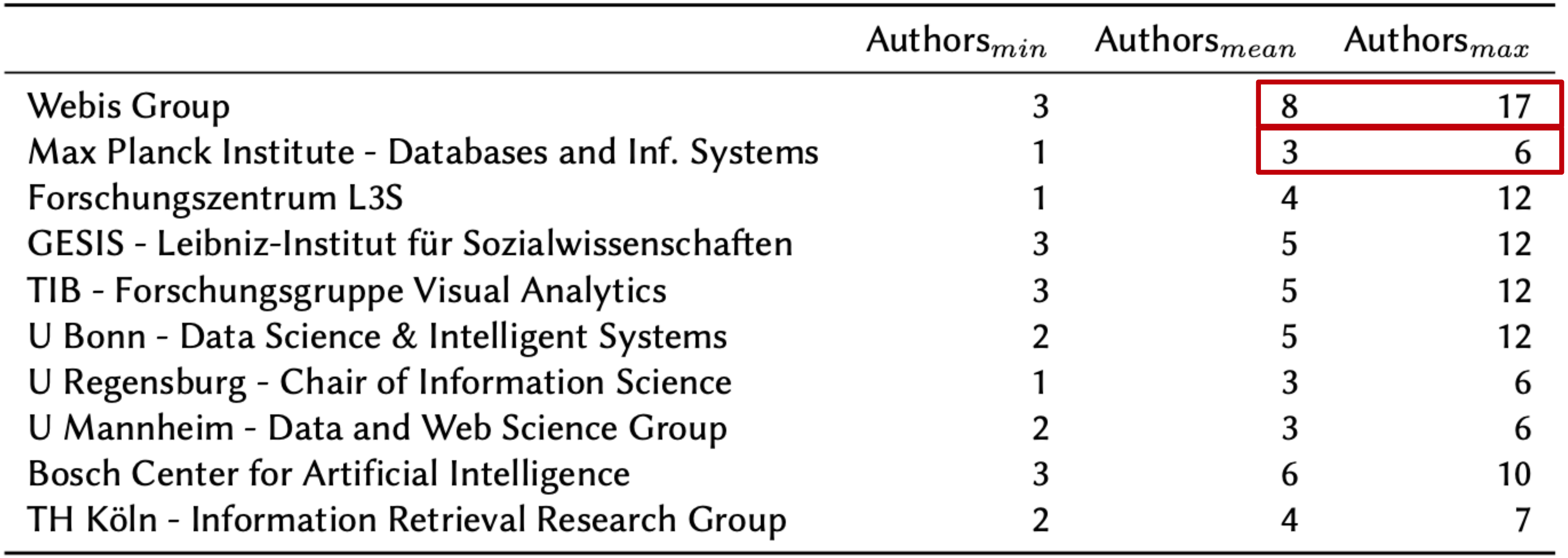
WebIS and L3S as “virtual groups” that are often co-affiliated with universities or other research groups

- Splitting up those large groups wouldn’t have changed a lot…

Not only universities

- one commercial research institute (Bosch Center for AI),
- one university of applied sciences (TH Köln), and
- two non-university research centers (GESIS and TIB)

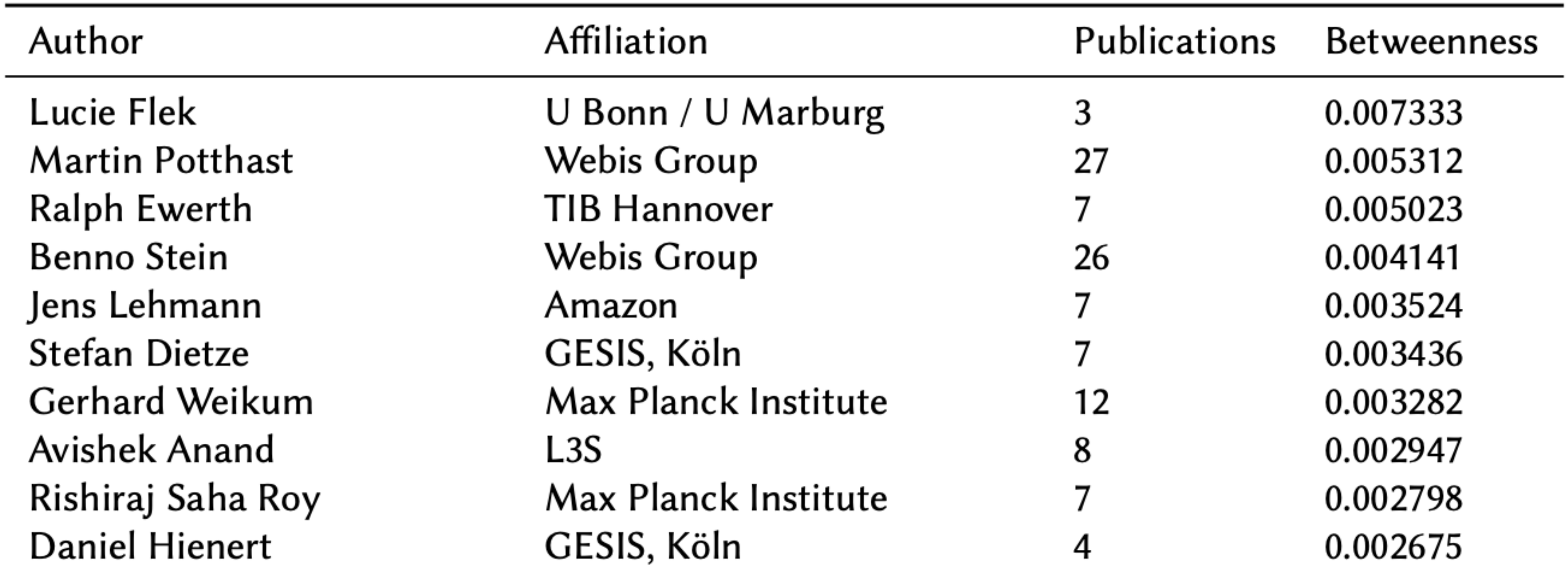
Co-authorships for the top 10 IR groups



	Authors_{min}	Authors_{mean}	Authors_{max}
Webis Group	3	8	17
Max Planck Institute - Databases and Inf. Systems	1	3	6
Forschungszentrum L3S	1	4	12
GESIS - Leibniz-Institut für Sozialwissenschaften	3	5	12
TIB - Forschungsgruppe Visual Analytics	3	5	12
U Bonn - Data Science & Intelligent Systems	2	5	12
U Regensburg - Chair of Information Science	1	3	6
U Mannheim - Data and Web Science Group	2	3	6
Bosch Center for Artificial Intelligence	3	6	10
TH Köln - Information Retrieval Research Group	2	4	7

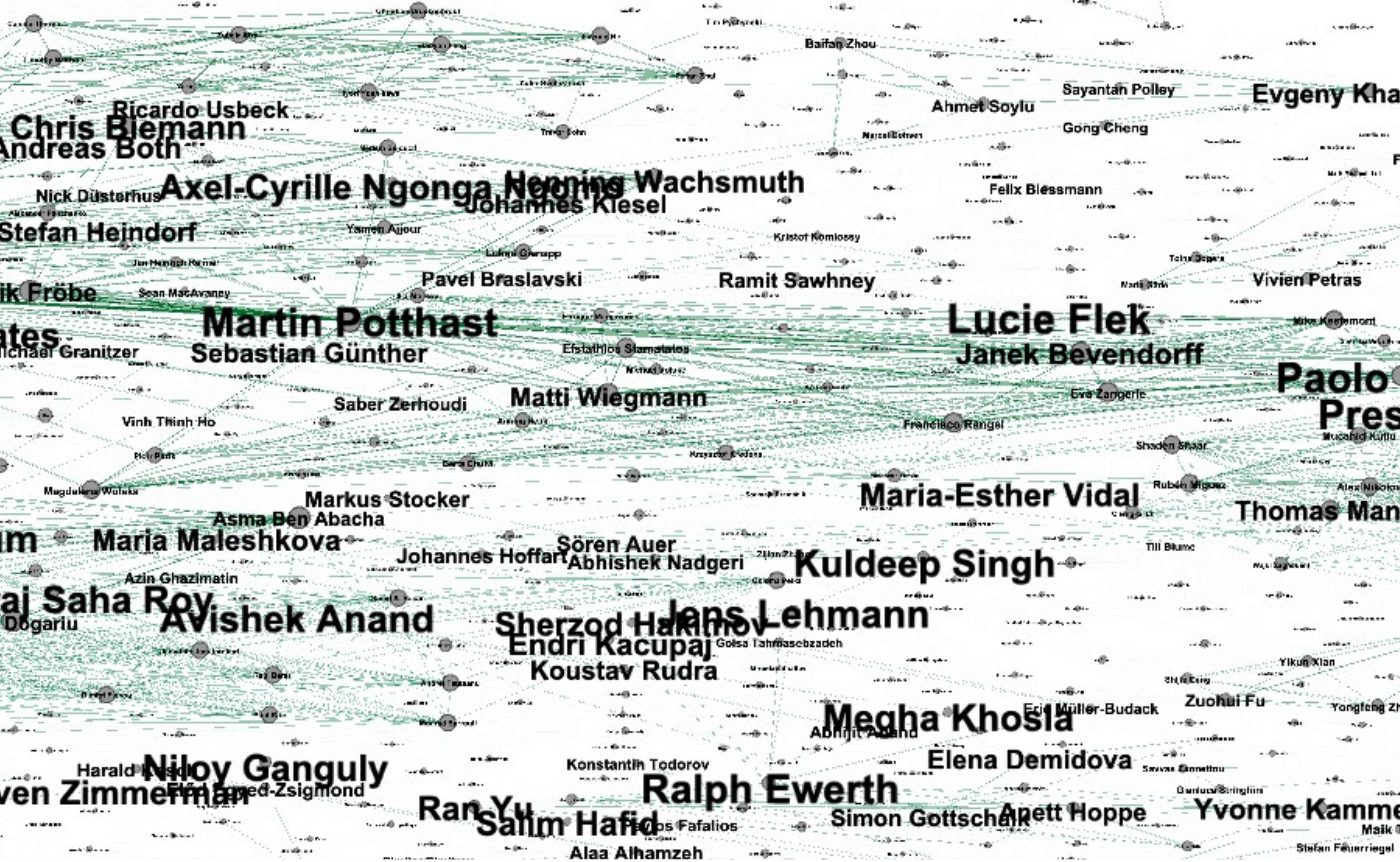
- On average, the top ten groups published papers with 4.83 authors, while on the whole data set the average number of authors was 4.98.
- Webis vs Max Planck – highest vs. lowest number of authors per paper
- The number of authors alone can therefore not explain the publication success of a group.

Co-author network



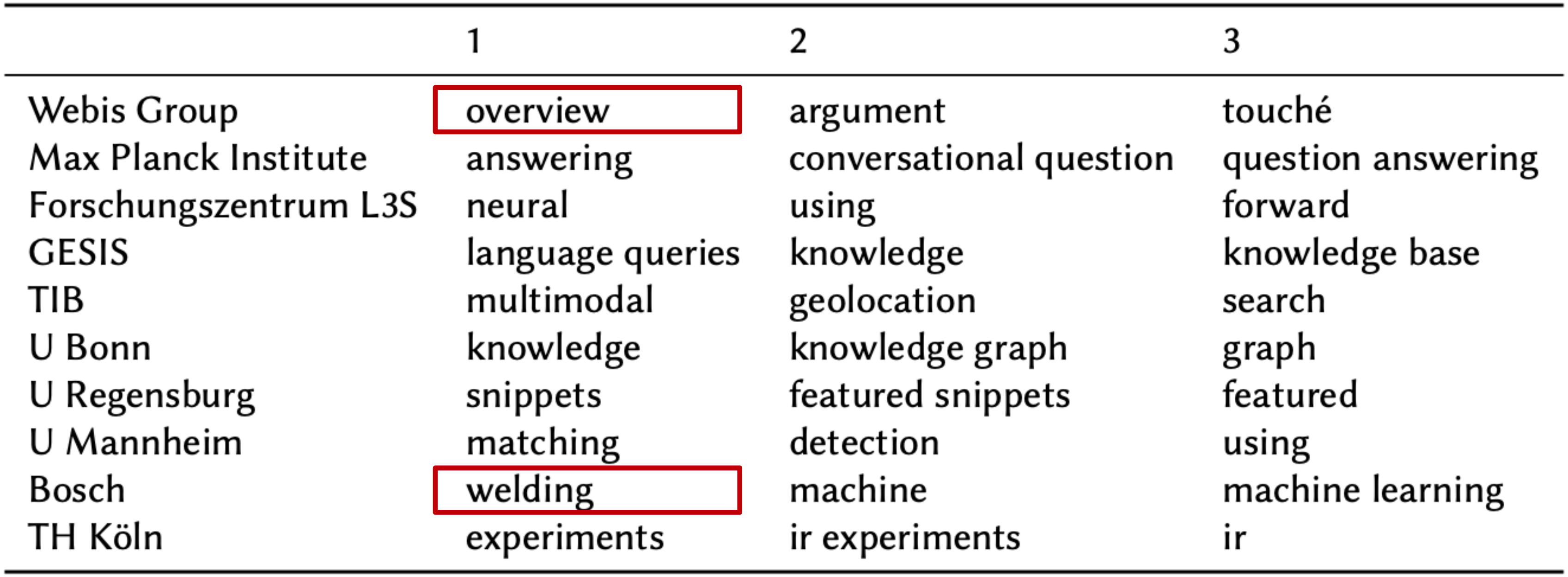
Author	Affiliation	Publications	Betweenness
Lucie Flek	U Bonn / U Marburg	3	0.007333
Martin Potthast	Webis Group	27	0.005312
Ralph Ewerth	TIB Hannover	7	0.005023
Benno Stein	Webis Group	26	0.004141
Jens Lehmann	Amazon	7	0.003524
Stefan Dietze	GESIS, Köln	7	0.003436
Gerhard Weikum	Max Planck Institute	12	0.003282
Avishek Anand	L3S	8	0.002947
Rishiraj Saha Roy	Max Planck Institute	7	0.002798
Daniel Hienert	GESIS, Köln	4	0.002675

- All author collaborations form a network of 1159 nodes (authors) and 4907 edges (co-authorship relations)
- Lucie Flek (U Bonn and U Marburg), with only three publications in total, but these papers were published at WWW, SIGIR, and ECIR and had no single overlap in co-authors



Get the GEPHI files:
https://github.com/irgroup/LWDA2023-IR-community

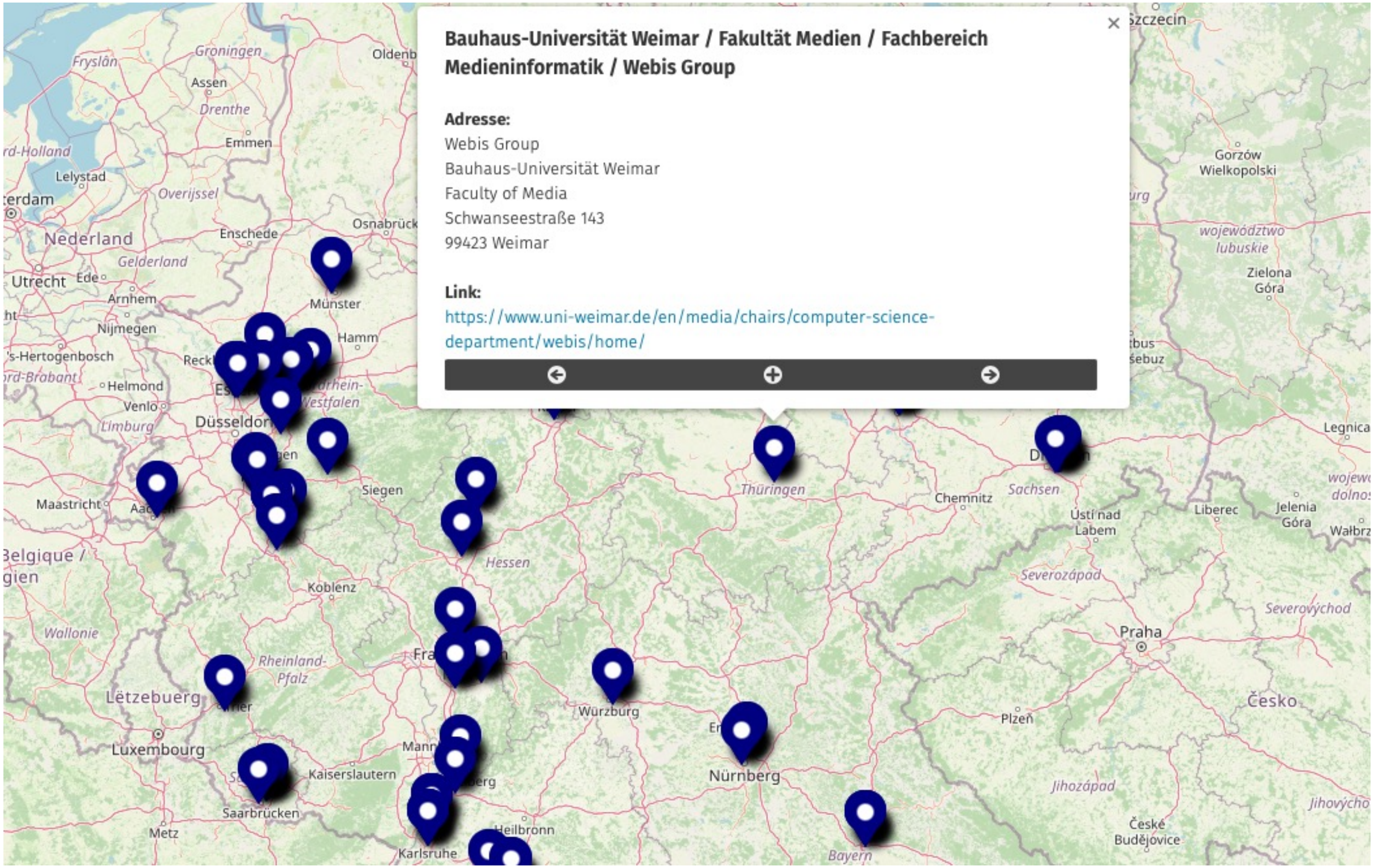
Topics of publications per group



	1	2	3
Webis Group	overview	argument	touché
Max Planck Institute	answering	conversational question	question answering
Forschungszentrum L3S	neural	using	forward
GESIS	language queries	knowledge	knowledge base
TIB	multimodal	geolocation	search
U Bonn	knowledge	knowledge graph	graph
U Regensburg	snippets	featured snippets	featured
U Mannheim	matching	detection	using
Bosch	welding	machine	machine learning
TH Köln	experiments	ir experiments	ir

- Top 3 TF-IDF terms/bi-grams from publication titles

Visualizing IR research groups



Bauhaus-Universität Weimar / Fakultät Medien / Fachbereich Medieninformatik / Webis Group

Adresse:
Webis Group
Bauhaus-Universität Weimar
Faculty of Media
Schwanseestraße 143
99423 Weimar

Link:
https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/home/

http://u.osmfr.org/m/931946/

Limitations and next steps

Small-scale scientometric study of the German IR community using publications from 2020 till mid-2023, with limitations:

- time frame in the middle of the COVID-19 pandemic and too short
- six IR-relevant conference, but missing some like JCDL, ICTIR
- CIKM might have introduced some shift, leaving out TREC and CLEF workshop proceedings left out many relevant papers

Next steps:

- extend to other conferences, longer time span, fine-grained affilitation / topical filtering
- invite underrepresented groups to FGIR

https://github.com/irgroup/LWDA2023-IR-community

References

- [1] Y. Ding, G. G. Chowdhury, S. Foo, Bibliometric cartography of information retrieval re- search by using co-word analysis, Information Processing & Management 37 (2001) 817–842. URL: https://www.sciencedirect.com/science/article/pii/S0306457300000510. doi:https://doi.org/10.1016/S0306-4573(00)00051-0.
- [2] M. Baumgartner, C. Schlögl, Die tagungsbände des internationalen symposiums für informationswissenschaft in szientometrischer analyse, in: A. Osswald, M. Stempfhuber, C. Wolff (Eds.), "Open Innovation" - Neue Perspektiven im Kontext von Information und Wissen: 10. Internationalen Symposiums für Informationswissenschaft, ISI 2007, Köln, Germany, 30. Mai - 1. Juni 2007, volume 46 of Schriften zur Informationswissenschaft, UVK, 2007, pp. 43–59. URL: https://doi.org/10.5281/zenodo.4134714. doi:10.5281/zenodo. 4134714.
- [3] D. Lewandowski, S. Haustein, What does the german-language information science community cite? - an analysis of the german information science handbook "grundlagen der praktischen information und dokumentation", in: F. Pehar, C. Schlögl, C. Wolff (Eds.), Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science, ISI 2015, Zadar, Croatia, May 19-21, 2015, volume 66 of Schriften zur Informationswissenschaft, Verlag Werner Hülsbusch, 2015, pp. 93–104. URL: https://doi.org/10.5281/zenodo.17973. doi:10.5281/zenodo.17973.
- [4] V. Larivière, C. R. Sugimoto, B. Cronin, A bibliometric chronicling of library and information science's first hundred years, Journal of the American Society for Information Science and Technology 63 (2012) 997–1016. URL: https://onlinelibrary. wiley.com/doi/abs/10.1002/asi.22645. doi:https://doi.org/10.1002/asi.22645. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22645.
- [5] C. V. Thornley, S. J. McLoughlin, A. C. Johnson, A. F. Smeaton, A bibliometric study of Video Retrieval Evaluation Benchmarking (TRECVid): A methodological analysis, Journal of Information Science 37 (2011) 577–593. URL: http://journals.sagepub.com/doi/10.1177/ 0165551511420032. doi:10.1177/0165551511420032.
- [6] B. Larsen, The Scholarly Impact of CLEF 2010–2017: A Google Scholar Analysis of CLEF Proceedings and Working Notes, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, volume 41, Springer International Publishing, Cham, 2019, pp. 547–554. URL: http://link.springer.com/10.1007/978-3-030-22948-1_22. doi:10. 1007/978-3-030-22948-1_22, series Title: The Information Retrieval Series.

References (contd.)

- [7] É. Archambault, É. Vignola-Gagné, G. Côté, V. Larivière, Y. Gingrasb, Benchmarking scientific output in the social sciences and humanities: The limits of existing databases, Scientometrics 68 (2006) 329–342. URL: http://link.springer.com/10.1007/s11192-006-0115-z. doi:10.1007/s11192-006-0115-z.
- [8] C. Michels, M. Neumann, P. Schaer, R. Schenkel, Conference indexing in digital libraries: A ranking model and case study on dblp, in: Proceedings of the 10th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 42nd European Conference on Information Retrieval, BIR@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only], 2020, pp. 30–41. URL: http://ceur-ws.org/Vol-2591/paper-04.pdf.
- [9] I. Masic, S. M. Jankovic, Inflated Co-authorship Introduces Bias to Current Scientometric Indices, Medical Archives 75 (2021) 248–255. URL: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC8563053/. doi:10.5455/medarh.2021.75.248-255.
- [10] Z. Zheng, B. Zhou, D. Zhou, A. Soylu, E. Kharlamov, Executable knowledge graph for transparent machine learning in welding monitoring at bosch, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 5102–5103. URL: https://doi.org/10.1145/3511808.3557512. doi:10.1145/3511808.3557512.
- [11] B. Zhou, Y. Svetashova, S. Byeon, T. Pychynski, R. Mikut, E. Kharlamov, Predicting quality of automated welding with machine learning and semantics: A bosch case study, in: M. d’Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 2933–2940. URL: https://doi.org/10. 1145/3340531.3412737. doi:10.1145/3340531.3412737.