On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch

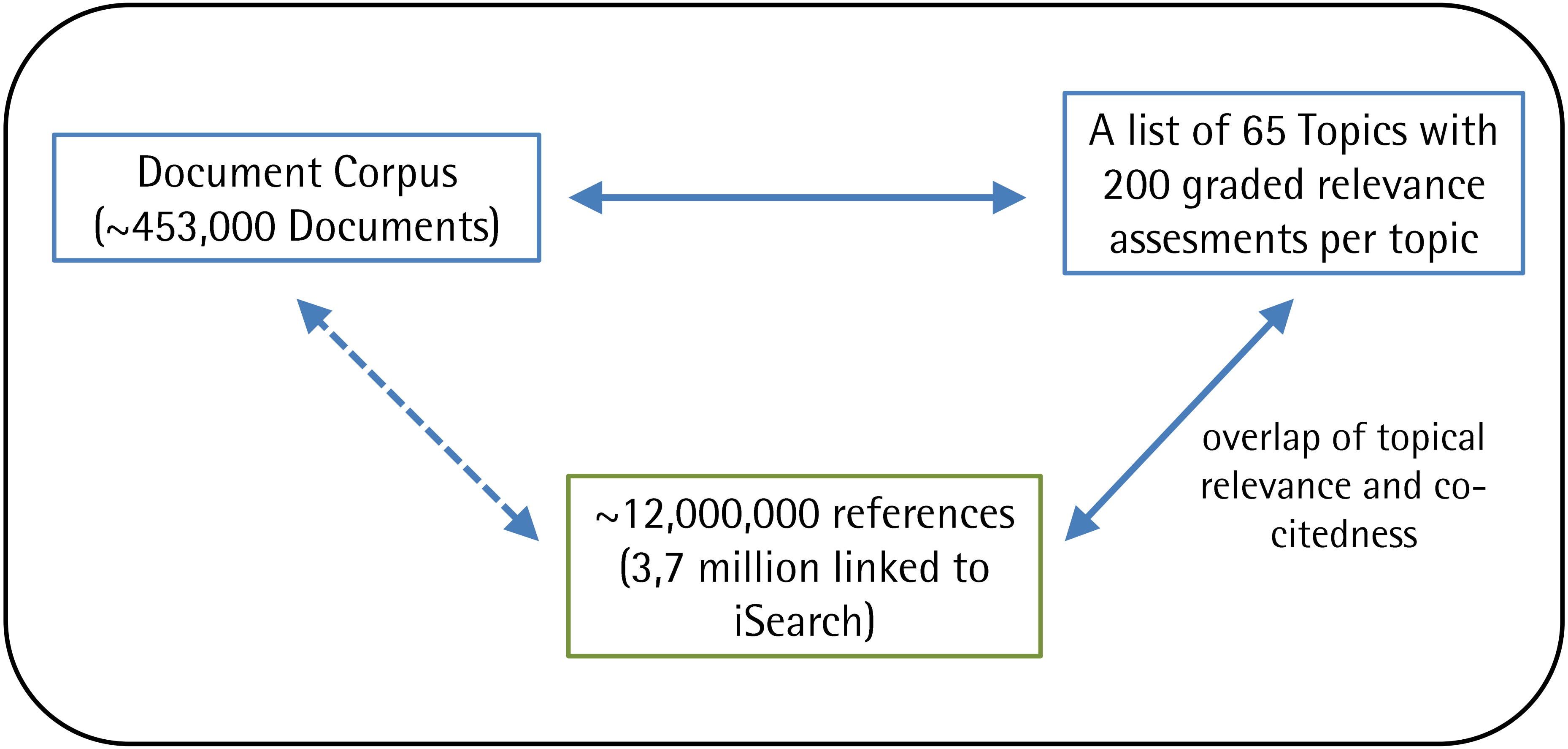
Workshop on Bibliometric-enhanced Information Retrieval
Amsterdam, 2014-04-13

Zeljko Carevic and Philipp Schaer
firstname.lastname@gesis.org

Motivation

- Bibliometric-enhanced Information Retrieval
  - One interpretation: Using citation data to improve retrieval
  - Could not be evaluated by IR standards up to now
  - Test collection now available: iSearch (Lykke et al. 2010)
- → Pretest of co-citation analysis using iSearch

iSearch: Possibilities



Document Corpus (~453,000 Documents)

A list of 65 Topics with 200 graded relevance assesments per topic

~12,000,000 references (3,7 million linked to iSearch)

overlap of topical relevance and co-citedness

Research Questions

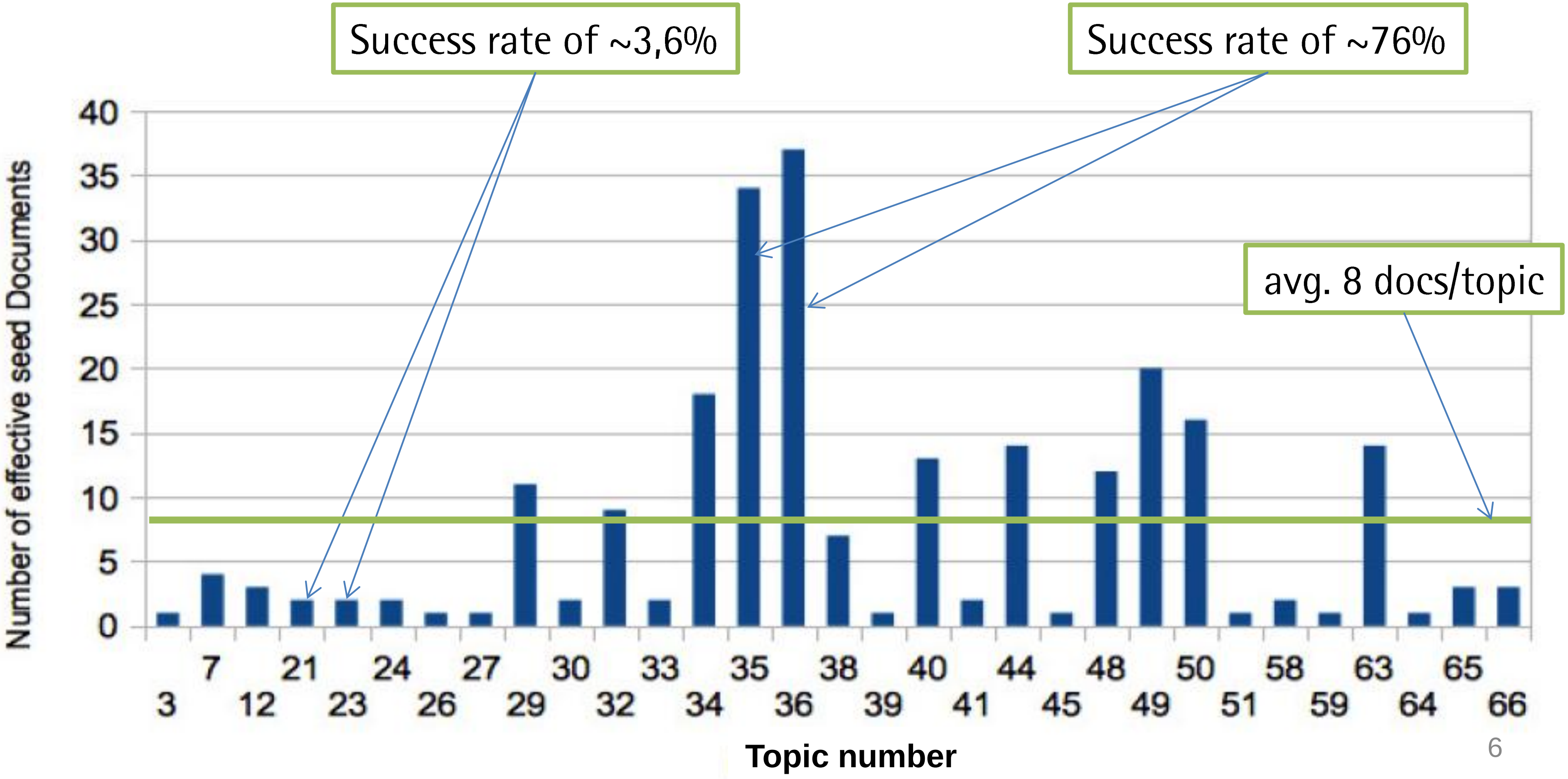
1. **Is the iSearch Test-Collection suitable for Co-Citation analysis?**

2. Can we find an overlap between documents relevant to a given topic and the results of a Co-Citation analysis to the given topic?

Co-citation analysis with iSearch

- Technically:
  - 1,6 million references with internal IDs
  - 3,4 million references without IDs (author/venue/year)
- Information coverage:
  - Sparse coverage per topic ...

Number of seed documents with at least one potential candidate per topic



Success rate of ~3,6%

Success rate of ~76%

avg. 8 docs/topic

Research Questions

1. Is the iSearch Test-Collection suitable for Co-Citation analysis?

2. **Can we find an overlap between documents relevant to a given topic and the results of a Co-Citation analysis to the given topic?**
    1. Due to the sparseness we could only ...

How to rank Results of a Co-Citation Analysis?

Ranking co-cited documents with TF*IDF (White 2010)

IR – TF*IDF ranking

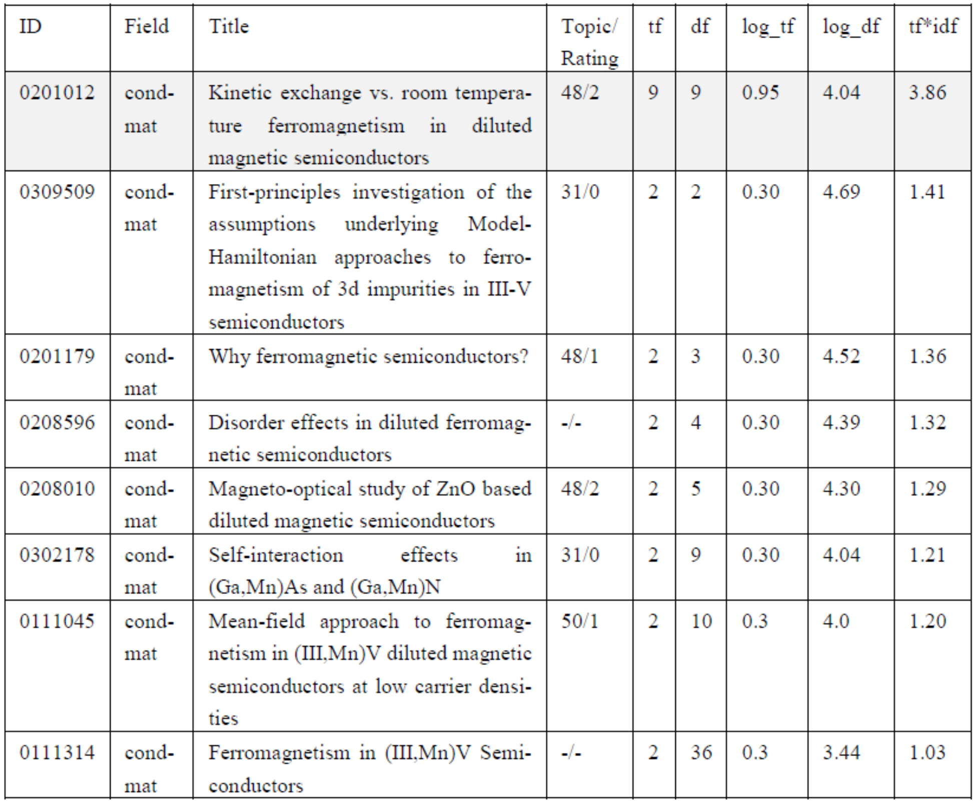
- Starts with a query term
- **tf** = Term frequency in current doc
- **df** = Number of docs query term apears in
- **TF*IDF** = similarity between **doc** and **query term**

Co-Citation – TF*IDF ranking

- Start with a seed doc
- **tf** = Number of times a doc is co-cited
- **df =** Number of times a doc is cited in the corpus overall
- **TF*IDF** = similarity between **doc** and the **seed**

Example Result for Topic 48

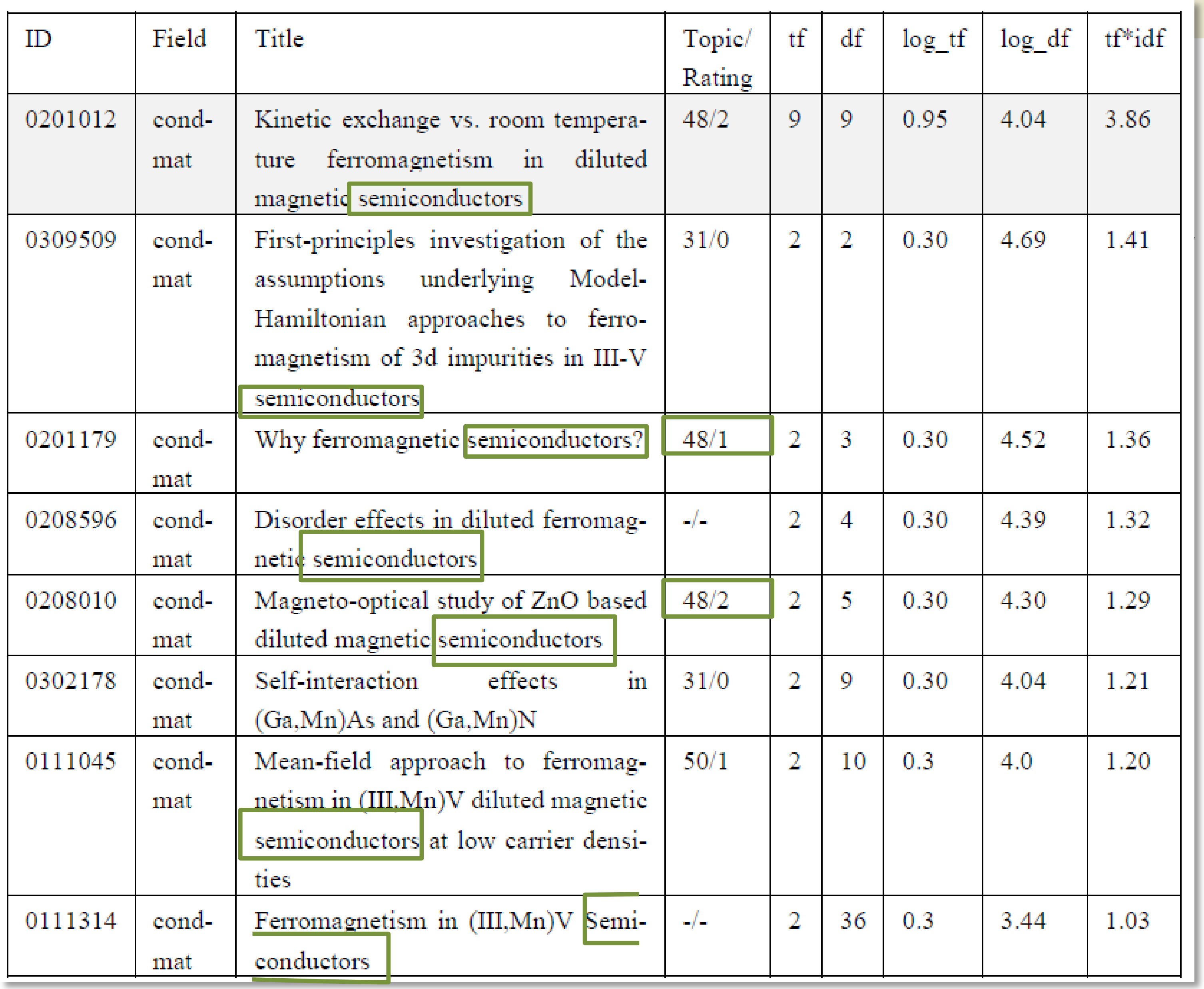
Seed document: *Kinetic exchange vs. Room temperature ferromagnetism in diluted magnetic semiconductors. Rated* ***fairly relevant*** *to the given Topic*



ID	Field	Title	Topic/ Rating	tf	df	log_tf	log_df	tf*idf
0201012	cond-mat	Kinetic exchange vs. room temperature ferromagnetism in diluted magnetic semiconductors	48/2	9	9	0.95	4.04	3.86
0309509	cond-mat	First-principles investigation of the assumptions underlying Model-Hamiltonian approaches to ferromagnetism of 3d impurities in III-V semiconductors	31/0	2	2	0.30	4.69	1.41
0201179	cond-mat	Why ferromagnetic semiconductors?	48/1	2	3	0.30	4.52	1.36
0208596	cond-mat	Disorder effects in diluted ferromagnetic semiconductors	-/-	2	4	0.30	4.39	1.32
0208010	cond-mat	Magneto-optical study of ZnO based diluted magnetic semiconductors	48/2	2	5	0.30	4.30	1.29
0302178	cond-mat	Self-interaction effects in (Ga,Mn)As and (Ga,Mn)N	31/0	2	9	0.30	4.04	1.21
0111045	cond-mat	Mean-field approach to ferromagnetism in (III,Mn)V diluted magnetic semiconductors at low carrier densities	50/1	2	10	0.3	4.0	1.20
0111314	cond-mat	Ferromagnetism in (III,Mn)V Semiconductors	-/-	2	36	0.3	3.44	1.03

Seed (

magn



ID	Field	Title	Topic/ Rating	tf	df	log_tf	log_df	tf*idf
0201012	cond-mat	Kinetic exchange vs. room temperature ferromagnetism in diluted magnetic semiconductors	48/2	9	9	0.95	4.04	3.86
0309509	cond-mat	First-principles investigation of the assumptions underlying Model-Hamiltonian approaches to ferromagnetism of 3d impurities in III-V semiconductors	31/0	2	2	0.30	4.69	1.41
0201179	cond-mat	Why ferromagnetic semiconductors?	48/1	2	3	0.30	4.52	1.36
0208596	cond-mat	Disorder effects in diluted ferromagnetic semiconductors	-/-	2	4	0.30	4.39	1.32
0208010	cond-mat	Magneto-optical study of ZnO based diluted magnetic semiconductors	48/2	2	5	0.30	4.30	1.29
0302178	cond-mat	Self-interaction effects in (Ga,Mn)As and (Ga,Mn)N	31/0	2	9	0.30	4.04	1.21
0111045	cond-mat	Mean-field approach to ferromagnetism in (III,Mn)V diluted magnetic semiconductors at low carrier densities	50/1	2	10	0.3	4.0	1.20
0111314	cond-mat	Ferromagnetism in (III,Mn)V Semiconductors	-/-	2	36	0.3	3.44	1.03

Discussion and future work

- Preliminary results of experiments using iSearch test collection.
- Using only internal reference identifiers did not retrieve a high enough number of documents.
- Expand the co-citation analysis by using:
  - Authors
  - Titles
  - Journal
  - Publication Year
- Implement citation analysis in an IR System for an evaluation of the recommended documents
- Source code available at: https://github.com/ZCarevic/iSearchCitationAnalysis

References

- [1] Buckley, C.: Why current IR engines fail. Inf. Retr. 12, 6, 652–665 (2009).
- [2] Lykke, M. et al.: Developing a Test Collection for the Evaluation of Integrated Search. In: Gurrin, C. et al. (eds.) Advances in Information Retrieval. pp. 627–630 Springer, Berlin, Heidelberg (2010).
- [3] White, H.: Some new tests of relevance theory in information science. Scientometrics. 83, 3, 653–667 (2010).