

# A Science Model Driven Retrieval Prototype

*Philipp Mayr, Philipp Schaer, Peter Mutschke*

*Abstract: This paper is about a better understanding of the structure and dynamics of science and the usage of these insights for compensating the typical problems that arises in metadata-driven Digital Libraries. Three science model driven retrieval services are presented: co-word analysis based query expansion, re-ranking via Bradfordizing and author centrality. The services are evaluated with relevance assessments from which two important implications emerge: (1) precision values of the retrieval service are the same or better than the tf-idf retrieval baseline and (2) each service retrieved a disjoint set of documents. The different services each favor quite other – but still relevant – documents than pure term-frequency based rankings. The proposed models and derived retrieval services therefore open up new viewpoints on the scientific knowledge space and provide an alternative framework to structure scholarly information systems.*

## 1. Introduction

In typical metadata-driven Digital Libraries three major difficulties arise: (1) the vagueness between search and indexing terms, (2) the information overload by the amount of result records obtained by information retrieval (IR) systems, and (3) the problem that pure term frequency based rankings, such as term frequency – inverse document frequency (*tf-idf*), provide results that often do not meet user needs.<sup>1</sup>

We will present an overall approach to use computational science models as enhanced search stratagems<sup>2</sup> within a scholarly IR environment. These computational models can be implemented within scholarly information portals. We assume that a user's search should improve by using these science model driven search tactics when interacting with a scientific information system.

This paper will at first introduce three scientific models: (1) co-word analysis and the derived concept of search term recommendation, (2) coreness of journals and (3) centrality of authors. The basic assumptions and concepts are presented.<sup>3</sup> The section on implementation deals with the set up prototype system that operationalizes the three models. The conducted evaluation and study with 73 students is described in the following section.

---

<sup>1</sup> Mayr et al. 2008.

<sup>2</sup> Bates 1990.

<sup>3</sup> See in detail Mayr et al. 2011.

The paper closes with a discussion of the observed results and the conclusion on the presented models.

## 2. *Models for Information Retrieval Enhancement*

The standard model of IR is the *tf-idf* model which proposes a text-based relevance ranking.<sup>4</sup> As *tf-idf* is text-based it assigns a weight to term  $t$  in document  $d$  which is influenced by different occurrences of  $t$  and  $d$ . Variations of the basis term weighing process have been proposed, like normalization of document length or by scaling the *tf* values but the basic assumption stays the same.

### 2.1 *Query Expansion via Search Term Recommendation*

Search Term Recommenders (STR) are an approach to compensate the long known language problem in Information Retrieval<sup>5</sup>: Searching an information system a user has to come up with the “correct” query terms so that they best match the document language to get an appropriate result.

STR are based on statistical co-word analysis and build associations between free terms (i.e. from title or abstract) and controlled terms (i.e. from a thesaurus) which are used during a professional indexation of the documents (see Figure 1). The co-word analysis implies a semantic association between the free and the controlled terms. The more often terms co-occur in the text the more likely it is that they share a semantic relation. These relations can be calculated and operationalized by different algorithms like LSA, PLSA, SVM and many others. In our setup we use STR for automatic query expansion where the original query of the user is enhanced with semantical “near” terms from a controlled vocabulary.

### 2.2 *Bradfordizing*

Bradfordizing is an alternative mechanism to re-rank result lists according to core journals to bypass the problem of very large and unstructured result sets (see Figure 2). The approach of Bradfordizing is to use characteristic concentration effects (Bradford’s law of scattering) that appear typically in journal literature. Bradfordizing defines different zones of documents which are based on the frequency counts in a given document set. Documents in core journals – journals which publish frequently on a topic – are

---

<sup>4</sup> Manning et al. 2008.

<sup>5</sup> Blair 2003; Petras 2006.

ranked higher than documents which were published in journals from the following zones. In Information Retrieval a positive effect on the search result can be assumed in favor of documents from core journals.<sup>6</sup>

The screenshot shows the 'Information Retrieval Value-added Services' interface. At the top, there is a navigation bar with 'Home', 'Applikationslabor', 'Prototypen', and 'Kontakt'. Below this, the search area contains a query box with 'luhmann' and a 'Search' button. A checkbox for 'Only show metadata sets which include an abstract' is present. The 'Automatic Query Expansion' section shows 'Thesaurus Sozialwissenschaften (de)' and 'Default relevance ranking'. The expanded query is: '[Systemtheorie, Autopoiesis, Kontingenz, Luhmann, N.]'. The results list shows 6410 hits, with the first two items being: 1. 'Droit et societe dans la sociologie de Niklas Luhmann : fondes en contingence (1997)' and 2. 'Soziologie der Behinderung und soziologische Systemtheorie - kritische Anmerkungen zur Systemtheorie von Niklas Luhmann und ihrer Rezeption in der Behindertenpädagogik (2004)'. On the right, the 'Interactive query enhancement' section shows a 'term cloud' with terms like 'Kommunikationsmedien', 'Individualität', 'verstehende Soziologie', 'soziales System', 'Theorievergleich', 'soziologische Theorie', 'Kommunikation', 'Gesellschaftstheorie', 'Komplexität', 'Kontingenz', 'Luhmann, N.', 'Autopoiesis', 'Sinn', 'soziale Differenzierung', 'System Soziologie', 'Erkenntnistheorie', 'Selbststeuerung', and 'Systemtheorie'.

Figure 1: Mapping between user terms and controlled terms in the IRM prototype. Example search term “luhmann” and highly associated controlled terms (controlled context on the right) from a STR build on the Thesaurus Sozialwissenschaften.

## Information Retrieval Value-added Services

This screenshot shows the same search interface as Figure 1, but with the 'Automatic Query Expansion' set to 'No Query Expansion' and the 'Rerank the result list' set to 'Core Journals'. The search results are reranked based on core journals.

Total hits: 1077

1. 'The Final Form of Perliminarity'. Views from the Experience of Theory (2005) journalarticle in *Soziale Systeme* 2005, 11, 1, 14-31. ( 0948-423X ) by Stiegler, Bernd; Roesler, Alexander
2. 'The Simulation of Social Systems by Means of Systems Theoretical Mechanisms -- A Macro Simulation with Stella (2006) journalarticle in *Soziale Systeme* 2006, 12, 1, 157-195. ( 0948-423X ) by Daiker, Christian
3. 'The Forgotten Function of Forgetting: Revisiting Exploration and Exploitation in Organizational Learning (2006) journalarticle in *Soziale Systeme* 2006, 12, 1, 100-120. ( 0948-423X ) by Schoeneborn, Dennis; Blaschke, Steffen

## Interactive query enhancement

The 'Interactive query enhancement' section shows a 'term cloud' and 'search term suggestions'. Below that, the 'core journals' section lists several journals with their respective issue counts:

- Soziale Systeme (197)
- Zeitschrift für Soziologie (106)
- Soziologische Revue (56)
- Zeitschrift für Rechtssoziologie (50)
- Soziale Welt (48)
- Kölner Zeitschrift für Soziologie und Sozialpsychologie (39)
- Sociologia Internationalis (31)
- Österreichische Zeitschrift für Soziologie (21)
- Schweizerische Zeitschrift für Soziologie/Revue Suisse de Sociologie/Swiss Journal of

Figure 2: Using core journals for reranking in the IRM prototype. Search term “luhmann” and core journals (journal context on the right).

<sup>6</sup> White 1981; Mayr 2009.

### 2.3 Author Centrality

Author centrality is another way of re-ranking result sets. Here the concept of centrality in a network of authors is an additional approach for the problem of large and unstructured result sets. The intention behind this ranking model is to make use of knowledge about the interaction and cooperation behavior in special fields of research (see Figure 3). The (social) status and strategic position of a person in a scientific community is used too. The model is based on a network analytical view on a field of research and differs greatly from conventional text-oriented ranking methods like *tf-idf*.

A concrete criterion of relevance in this model is the centrality of authors from retrieved publications in a co-authorship network. The model calculates a co-authorship network based on the result set to a specific query. Centrality of each single author in this network is calculated by applying the betweenness measure and the documents in the result set are ranked according to the betweenness of their authors so that publications with very central authors are ranked higher in the result list.<sup>7</sup>

#### Information Retrieval Value-added Services

Query:

Only show metadata sets which include an abstract

**Automatic Query Expansion** **Rerank the result list**

Total hits: 3332

1. Code - Performances - Function: The Constitution of Societal Subsystems (1998) journalarticle in *Soziale Systeme* 1998, 4, 1, 175-183. by Schimank, Uwe
2. Code - Leistungen - Funktion : zur Konstitution gesellschaftlicher Teilsysteme (1998) journalarticle in *Soziale Systeme Zeitschrift für soziologische Theorie* Jg. 4 H. 1 ( 0948-423X ) by Schimank, Uwe
3. A political sociology for complex societies: Niklas Luhmann (2001) article in *Blackwell companions to sociology*The Blackwell companion to political sociology ( 0-631-21050-4 ) by Lange, Stefan; Schimank, Uwe

#### Interactive query enhancement

term cloud ▼

search term suggestions ▼

core journals ▼

central authors ▼

- Schimank, Uwe (0.0456667)
- Luhmann, Niklas (0.0440176)
- Kneer, Georg (0.0171039)
- Luhmann, Hans-Jochen (0.017)
- Nassehi, Armin (0.0101904)
- Tyrell, Hartmann (0.00826002)
- Greshoff, Rainer (0.00813967)
- Kron, Thomas (0.00769569)
- Japp, Klaus P. (0.00705069)
- Baecker, Dirk (0.00686217)

Figure 3: Using author centrality for reranking in the IRM prototype. Search term “luhmann” and central authors (author context on the right).

## 3. Implementation

All proposed models are implemented in a live information system using (1) the Solr search engine, (2) Grails Web framework and (3) Recommind Mindserver to demonstrate the general feasibility of the approaches. Solr is

<sup>7</sup> Mutschke 2004.

an open source search platform from the Apache Lucene project<sup>8</sup> which uses a *tfidf* based ranking mechanism<sup>9</sup>. The Mindserver is a commercial text categorization tools which was used to generate the STR. Both Bradfordizing and author centrality as re-rank mechanism are implemented as plugins to the open source web framework Grails. Grails is the glue to combine the different modules and to offer an interactive web-based prototype<sup>10</sup>.

These retrieval services can be applied in different query phases. In a typical search scenario a user first formulates his/her query, which can then be enriched by a STR that adds controlled descriptors from the corresponding document language to the query. With this new query a search in a database can be triggered. The search returns a result set which can be re-ranked using either Bradfordizing or author centrality. Since search is an iterative procedure this workflow can be repeated many times till the expected result set is retrieved.

#### 4. Evaluation

By measuring the contribution of our services to retrieval performance we expect deeper insights in the structure and the functions of the science system. The evaluation plays the role of a “litmus test” for the adequacy of the science models proposed.

##### 4.1 Methods

The standard approach to evaluate Information Retrieval systems is to do relevance assessments. In respect to a defined information need documents are marked as relevant or not relevant. Large standard test collections (like TREC, CLEF etc.) contain pre-assessed documents from domain experts where all containing documents are judged relevant or not relevant. Since modern collections usually are too large to be assessed in total only subsets of the collection are assessed. *Pooling* is used to disguise the origin of the document.<sup>11</sup> In this standard approach subsets of the collections are formed by pooling the top *n* documents returned by the different IR systems to be evaluated. In the next step the assessors have to judge the documents in the subsets without knowing the originating IR systems.

---

<sup>8</sup> Cf. <http://lucene.apache.org/Solr/>.

<sup>9</sup> Cf. [http://lucene.apache.org/java/2\\_4\\_0/scoring.html](http://lucene.apache.org/java/2_4_0/scoring.html).

<sup>10</sup> Cf. <http://www.gesis.org/beta/prototypen/irm/>.

<sup>11</sup> Voorhees and Harman 2005

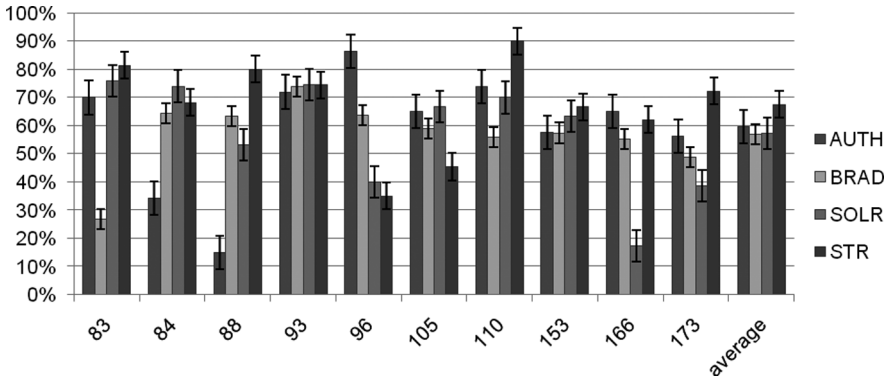


Figure 4: Precision for each topic and service (Relevance assessments per topic / total amount of single assessments), including standard error

## 4.2 Assessed Data Set

We conducted a user assessment with 73 students who used the SOLIS database with 369,397 single documents (including title, abstract, controlled keyword etc.). After a briefing each student had to choose one out of 10 different predefined topics. Topic title and the description were presented to form the information need. The pool was formed out of the top  $n=10$  ranked documents from each service and the initial *tf-idf* ranked result set respectively. Duplicates were removed, so that the size of the sample pools was between 34 and 39 documents each. The assessors could choose to judge relevant or not relevant (binary decision) – in case they didn't assess a document this document is ignored in later calculations.

The assessors did 43.78 assessments in average which sums up to 3,196 single relevance judgments in total. Only 5 participants did not fill in the assessment form completely, but 13 did more than one. Since every assessor could freely choose from the topics the assessments are not distributed evenly. Topic 83 was picked 16 times – topic 96 twice.

## 5. Results

### 5.1 Precision

The precision  $P$  of each service was calculated by

$$P = \frac{|r|}{|r + nr|}$$

for each topic, where  $|r|$  is the number of all relevant assessed documents and  $|r+nr|$  is the number of all assessed documents (relevant and not relevant). All precision values and numbers of relevance assessments can be seen in table 1. A graph of all precision values including standard error can be seen in Figure 4.

The average precision of the STR was highest (68%) compared to the baseline from the SOLR system (57%). The two alternative ranking methods Bradfordizing (BRAD) and author centrality (AUTH) scored 57% and 60% respectively.

## 5.2 *Inter-rater agreement*

The assessors in this experiment were not professionals and/or domain experts but mainly library and information science students. However Al-Maskari et al.<sup>12</sup> showed that this evaluation setup can produce relevant data. They compared official TREC with non-TREC assessors. The agreement changed due to the different topics and the actual ranking position of the assessed document and was between 75% and 82%. The agreement rates of these studies can be compared to our assessments where the overall agreement between all topics and all participants was 82%. 124 of 363 cases were perfect matches where all assessors agreed 100% (all relevant and non relevant judgments matched).

To further rate the reliability and consistency of agreement between the different assessments a statistical measure is needed. Fleiss' Kappa is a measure of inter-grader reliability for nominal or binary ratings.<sup>13</sup> It is an extension of Cohen's Kappa for two-grader ratings. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly.

Kappa scores can range from  $<0$  (less than chance) over 0.0 (chance) to 1.0 (full agreement). Different interpretations of Fleiss's Kappa were proposed.<sup>14</sup> All Kappa scores in our experiment range between 0.20-0.40 (topics 84, 110 and 153) and 0.40-0.52 (rest of topics) respectively. Average score was 0.4 and median score was 0.43 which are fair up to moderate levels of agreement or mainly acceptable in the more conservative interpretations.

---

<sup>12</sup> Al-Maskari et al. 2008.

<sup>13</sup> Fleiss 1971.

<sup>14</sup> Landis and Koch 1977 or Osman et al. 2010.

id	non relevant				relevant				precision (in %)			
	AUTH	BRAD	SOLR	STR	AUTH	BRAD	SOLR	STR	AUTH	BRAD	SOLR	STR
83	42	104	36	25	98	38	114	109	70,00	26,76	76,00	81,34
84	71	38	27	26	37	69	77	56	34,26	64,49	74,04	68,29
88	51	22	28	12	9	38	32	48	15,00	63,33	53,33	80,00
93	28	26	25	26	72	74	73	76	72,00	74,00	74,49	74,51
96	3	8	12	13	19	14	8	7	86,36	63,64	40,00	35,00
105	15	18	15	24	28	26	30	20	65,12	59,09	66,67	45,45
110	13	22	15	5	37	28	35	45	74,00	56,00	70,00	90,00
153	42	40	36	32	57	54	62	64	57,58	57,45	63,27	66,67
166	30	39	72	33	56	48	15	54	65,12	55,17	17,24	62,07
173	41	48	57	26	53	46	36	68	56,38	48,94	38,71	72,34
avg.									59,58	56,89	57,37	67,57

Table 1: Relevance judgments for each topic and service with calculated precision

### 5.3 *Overlap of top document result sets*

A comparison of the intersection of the relevant top 10 document result sets between each pair of retrieval service shows that the result sets are nearly disjoint. 400 documents (4 services \* 10 per service \* 10 topics) only had 36 intersections in total. Thus, there is no or very little overlap between the sets of relevant top ranked documents. The largest, but still very low overlap is among the standard *tf-idf* ranking from SOLR and STR which have 14 common documents.

## 6. *Discussion*

The evaluation of the STR enhanced system showed that term suggestions can provide an overview over different areas of discussion by adding new concepts. Additionally the STR can support different domains and perspectives.<sup>15</sup> This is known as a “query drift” in automatic query expansion but in the application of the STR the service retrieves more relevant documents. While the result set of a STR enhanced query grows and broadens (because of the OR-ing of the added terms) the first n=10 hits are more pre-

<sup>15</sup> See Petras 2006.



cise and narrowed. This contradiction can be explained with the high descriptive power of the controlled terms that are added to the query and is an indicator for the high quality of the semantic mapping between the language of scientific discourse (free text in title and abstract) and the language of documentation (controlled thesauri terms).

Discussing the results of the two proposed re-ranking methods Bradfordizing and author centrality brings up two central insights: (1) users get new result cutouts with other relevant documents which are not listed in the first section (first  $n=10$  documents) of the original list and (2) Bradfordizing and author centrality can be a helpful information service to positively influence the search process, especially for searchers who are new on a research topic and don't know the main publication sources or the central actors in a research field.

The Bradford approach always runs the risk and critic of disregarding important developments outside the core. It is often criticized to favor majority views and mainstream journals and ignores minority standpoints. This is a serious argument but it has to be seen as a problem of the data set producers because Bradfordizing only works with existing document sets, which are compiled (and prefiltered) by database producers.

The basic assumption of author centrality based ranking is that central authors are strongly associated with the mainstream topics of a research field and that the phenomenon of authors of high betweenness are supposed to be of high community driving relevance for the science system can be utilized. This perception of the strategic role of highly central actors in science might explain the high precision of rankings done by author betweenness: Authors of high betweenness address the key topics of a field.

## 7. Conclusion

Looking at the precision values and the overlap of the result sets two important insights can be noted: (1) precision values of the retrieval services are the same or better than the *tf-idf* retrieval baseline and (2) each service retrieved a disjoint set of documents. The different services each favor quite different – but still relevant – documents than pure term-frequency based rankings. The proposed models and derived  $r$  services open up new viewpoints on the scientific knowledge space and also provide an alternative framework to structure the science system.

In a next step we plan to evaluate the proposed retrieval services with strictly scientific topics and scientist who are assessing documents in their specific research field.

## *Acknowledgements*

Special thanks go to the students at Humboldt University (guided by Vivien Petras) and University of Applied Science in Darmstadt who took part in our assessment. We thank Hasan Bas who implemented the assessment tool.

This work was funded by DFG, grant no. INST 658/6-1.

## *References*

- Al-Maskari, Azzah; Sanderson, Mark; Clough, Paul. (2008). Relevance Judgments between TREC and Non-TREC Assessors. In: Proceedings of the 31<sup>st</sup> Annual International ACM SIGIR.
- Bates, Marcia J. (1990). Where Should the Person Stop and the Information Search Interface Start? In: *Information Processing & Management* 26 : 575-591.
- Blair, David C. (2003). Information Retrieval and the Philosophy of Language. In: *Annual Review of Information Science and Technology* 37 : 3-50.
- Fleiss, Joseph L. (1971). Measuring Nominal Scale Agreement Among Many Raters. In: *Psychological Bulletin* 76 (5) : 378-382.
- Landis, J. Richard; Koch, Gary. G. (1977). The Measurement of Observer Agreement for Categorical Data. In: *Biometrics* 33 (1) : 159-174.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Mayr, Philipp; Mutschke, Peter; Petras, Vivien. (2008). Reducing Semantic Complexity in Distributed Digital Libraries: Treatment of Term Vagueness and Document Re-ranking. In: *Library Review* 57 (3) : 213-224.
- Mayr, Philipp; Mutschke, Peter; Schaer, Philipp; Sure, York. (2011). Science Models as Value-Added Services for Scholarly Information Systems. In: *Scientometrics*.
- Mayr, Philipp. (2009). *Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken*. Dissertation. Humboldt-Universität zu Berlin, Berlin.
- Mutschke, Peter. (2004): Autorennetzwerke: Netzwerkanalyse als Mehrwertdienst für Informationssysteme. In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hrsg.). *Information zwischen Kultur und Marktwirtschaft: Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004)*, Chur, 6.-8. Oktober 2004. Konstanz: UVK Verl.-Ges. 141-162

- 
- Osman, Deanna; Yearwood, John; Vamplew, Peter. (2010). Automated Opinion Detection: Implications of the Level of Agreement between Human Raters. In: *Information Processing & Management* 46 (3) : 331–342.
- Petras, Vivien. (2006). *Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages*. Dissertation. Berkeley: University of California.
- Voorhees, Ellen M.; Harman, Donna K. (Eds.). (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge: The MIT Press.
- White, Howard D. (1981). 'Bradfordizing' Search Output: How It Would Help Online Users. In: *Online Review* 5 (1) : 47–54.