Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering

Muhammad Kamran Abbasi and Ingo Frommholz

Institute for Research in Applicable Computing
University of Bedfordshire
{muhammad.abbasi|ingo.frommholz}@beds.ac.uk

Abstract. In this paper we explore the potential of combining the principle of polyrepresentation with document clustering. Our idea is discussed and evaluated for polyrepresentation of information needs as wells as for document-based polyrepresentation where bibliographic information is used as representation. The main idea is to present the user with the highly ranked polyrepresentative clusters to support the search process. Our evaluation suggests that our approach is capable of increasing retrieval performance, but performance varies for queries with a high or low number of relevant documents.

1 Introduction

The key objective of information retrieval systems is to satisfy the user's information need in the provided context. In particular Interactive Information Retrieval (IIR) is supposed to support the user beyond just typing in queries. In this respect the principle of *polyrepresentation* is a highly recognized approach to IIR [1]. It suggests to use various information/data representations to integrate the context and interpretation of different actors into the information retrieval process. The representations can come from the same actor but for different purposes and are functionally different, or they can come from various actors and are cognitively different. Polyrepresentation may refer to the information space, the user's cognitive space or a combination thereof. In Bibliometrics, references citing another source establish a cognitively different representation of a document. Examples for different information need representations are work task descriptions and the query. The principle of polyrepresentation in IR could be described as follows: if multiple representations are pointing towards an information object it is more likely to be relevant to the user's information need. This is depicted in Figure 1. Let us assume $\mathcal{R}$ represents the relevance of a representation, hence $\mathcal{R}_1$ denotes the documents relevant to representation 1, $\mathcal{R}_2$ to representation 2 and so on, so the documents in $\mathcal{R}_1, \mathcal{R}_2$ and $\mathcal{R}_3$ are only relevant to these individual representations. The intersection of the two representations i.e, $\mathcal{R}_{12}$ $\mathcal{R}_{13}$ and $\mathcal{R}_{23}$ holds the documents relevant to the two respective representations, and the intersection of all three representation, $\mathcal{R}_{123}$ is the *total cognitive overlap*. According to the principle of polyrepresentation this set is supposed to hold the most relevant documents as evaluated in [2,3,4].

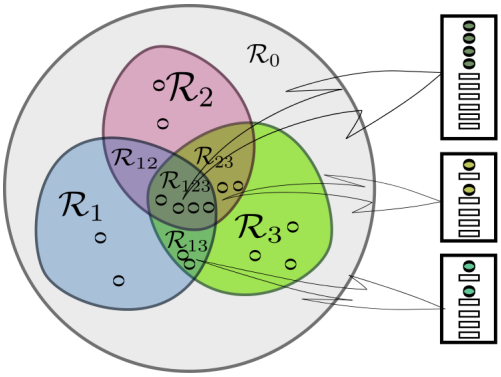

Fig. 1. Polyrepresentation and Clustering. The left hand side shows the relevance sets w.r.t. combinations of different representations. The right hand side shows the induced rankings. Small circles denote relevant documents.

The principle of polyrepresentation has been evaluated so far in ad hoc retrieval. However, while different combinations of representations were evaluated in a more static fashion, some open problems remain from a user perspective. A system may be able to combine different representations, but the system initially does not know whether and to what degree the user prefers some of them. A weighting mechanism as proposed in [5] may mitigate the situation, but still the system needs additional information. We argue that instead of presenting users with a ranked list of results, we may present them with clusters they can choose from. Clustering and polyrepresentation both create a partitioning of the information objects under consideration. One approach to combine document clustering and polyrepresentation is described by a "polyrepresentation cluster hypothesis": documents relevant to the same representation should appear in the same cluster [6].

The possible application of the above discussed approach could be presenting the clusters to the user in a way depicted in Figure 1. In this case, we present the user with a cluster containing a ranked list of documents representing the total cognitive overlap first. We assume the user browses the top k documents and then moves on to the next cluster (one where only a subset of all representations is relevant), depending on her representation preferences, where she examines again some top k documents.

2 Bibliometrics, Polyrepresentation and Document Clustering

The science modeling and the bibliometrics provides the means to analyze and quantify the structure and process of scholarly communication [7]. The methods range from citation analysis [8], co-citation clustering [9] to bradfordizing [10].

The developments in e-publishing and availability of the full text and meta-data regarding the scientific information objects increased the scope and applicability of the bibliometrics, hence more refined measures are needed [7]. The connections between IR, bibliometrics and relevance theory are discussed in [11]. The citation networks and clustering methods suitable for block modelling in similar context are discussed in [12]. The suitability of the bibliometirc measures for enhancement of retrieval in scholarly systems is presented and evaluated in [13,10]. The authors have evaluated bibliometrics approaches like bradfordizing for reranking and co-word model for query expansion in search term recommender and report performance improvement. The principle of polyrepresentation in similar context has been discussed in [14] using the references and citation information. We look at the principle of polyrepresentation as a method go along with the bibliometric approaches, because it allows the use of various representations, hence the bibliographic information i.e. authors, references and the citation context could be exploited as representations.

The Optimum Clustering Framework (OCF) [15] provides a theoretical justification for clustering and is based on the notion of using so-called query sets. We propose to apply the OCF for inferring the suitable candidate clusters representing the various polyrepresentation sets $\mathcal{R}$. The OCF operates on the probability of relevance $\Pr(R|d, q_i)$ of the document d with respect to a query q_i in the query set; since in polyrepresentation we are dealing with measuring the degree or probability of relevance for each representation, OCF is a suitable framework for our ideas. From this, each document is represented by the vector $\tau^T(d) = (\Pr(R|d, q_1), \ldots, \Pr(R|d, q_n))$ with n as the number of queries in the query set. The vectors are now used to cluster documents with the choice of clustering function depending on the overall setup. If each term in the collection is regarded as a query q_i the OCF simply models classical clustering using term-based similarity.

To set up the OCF for polyrepresentation we need to distinguish between polyrepresentation of information needs on the one hand and polyrepresentation of documents on the other hand. In order to apply clustering to information need polyrepresentation let REP_{in} be the representations of an information need in. $\Pr(R|d, r_i)$ is computed for each document d and $r_i \in REP_{in}$. From this we create a vector $\tau^T(d) = (\Pr(R|d, r_1), \ldots, \Pr(R|d, r_n))$ with $n = |REP_{in}|$. When applying polyrepresentation of documents, REP_d consists of the different representations rd_i of a document d. In our case we assume that the information need is represented by the query q alone[1]. We therefore need to compute $\Pr(R|rd_i, q)$ and we get $\tau^T(d) = (\Pr(R|rd_1, q), \ldots, \Pr(R|rd_n, q))$ with $n = |REP_d|$.

[1] The combination of document and information need polyrepresentation is subject to future work.

3 Evaluation

Collection and clustering approach In order to evaluate the proposed approach, the PF part of the iSearch[2] [16] collection is used. This sub collection contains full text articles related to the physics domain. The collection comes with 65 search tasks, each search task comprises upon five information need representations i.e. Search Terms, Work Task, Current Information Need, Ideal Answer and Background Knowledge. These five representation make REP_{in} for the $\mathcal{R}$ set for information need based polyrepresentation. In a first step to estimate $\Pr(R|d, r_i)$ BM25 based document weights were computed for every information need representation r_i using the Terrier IR platform [17]. The weights for each representation were combined using CombSum [18] to get a retrieval weight for each document based on the single representations. OCF-based clustering was performed with k-means. k (in k-means) was set to $2^{|REP_{in}|}$ to create exactly as many clusters as there are combinations of relevant representation in polyrepresentation.

The second exploration has been about document based polyrepresentation, for this the representations i.e title, abstract, body, context, and references were extracted from the documents. The context c of a document d has been extracted from all the articles cited in d, initially only the title and abstract of the cited document were combined and used as a context. Hence, the context for a document d becomes the concatenation of all abstracts and titles of the cited documents in it, as depicted in Figure 2. The derived representations in this

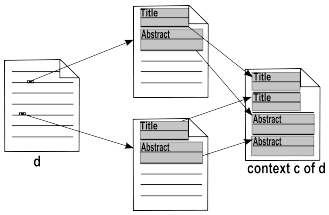

Fig. 2. Document Context

part make the REP_d part for the $\mathcal{R}$ set. The representation here has also been indexed and the BM25 weights were computed as discussed above. The *Search Term* part of the all 65 topics in the iSearch collection has been used as a query.

Baseline and Polyrepresentative Clustering The goal of our evaluation is to verify the effect of clustering on polyrepresentation. Therefore our baseline is to combine the BM25 weights mentioned above from each representation using CombSum, which creates a ranked list that utilises different representations but no clustering.

[2] http://itlb.dbit.dk/~isearch/

To compare our polyrepresentative clustering approach against the baseline we create a new cluster-based ranking. As mentioned before we assume a system that presents the user with the total cognitive overlap cluster first, i.e. the user is presented with a ranked list of documents in this cluster. We further assume that the user examines the top k documents and then moves on to a different cluster. To *simulate* this behaviour for evaluation purposes our strategy is as follows. We first create a clustering $\mathcal{C}$ as described above. From this we produce an artificial ranking as described in Algorithm 1. The ranking simulates the documents that a user would see along the path when examining the top k documents of each cluster. We rank the clusters in $\mathcal{C}$ so that the top ranked cluster is an approximation of the total cognitive overlap (please see [6] for a further discussion). We now process the cluster ranking in descending order. We rank the documents in each cluster and append the top$k$ documents to the ranking.

Require: Clustering $\mathcal{C}$, k
  $r \leftarrow ()$ {The ranking, initially an empty list}
  $C \leftarrow$ ranked list of clusters in $\mathcal{C}$ (using eF or SD)
  for all cluster $c \in C$ **do**
    $l \leftarrow$ ranked list of documents in c {process C in descending weight order}
    for $i = 1$ to k **do**
      $r \leftarrow r + l[i]$ {append document at rank i to r}
    end for
  end for
  return r

Algorithm 1: Cluster-based ranking

To create the ranking of the clusters, we have used two measures. One is the OCF based *expected F-measure (eF)* as described in [15]. The second is the sparsity density of the REP weight matrix constituting the cluster C, with the idea to identify the clusters where many or all representations have contributed and the corresponding cluster point matrix is less or not sparse. The sparsity density (SD) of the cluster matrix is computed by counting the non zero elements in the cluster matrix divided by the size of the matrix. In this approach the matrix size is equal to the number of documents in the cluster (rows) multiplied with the number of $REPs$ (columns). The range of the density of the cluster is between 0 and 1 where values closer to 1 show higher density. We have used eF and SD to rank the clusters. The top k documents then were extracted form each cluster and merged to create the rank for the **trec_eval**, as described in the Algorithm 1. For comparing the baseline and proposed approach we used $k = 5$ and $k = 10$. We extracted binary relevance judgements from the grades iSearch ones with a value > 1 meaning relevance.

Table 1. Queries with high number of relevant documents

runid	IN BM25 High	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*	DOC BM25 High	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*
num_q	19	19	19	19	19	19	19	19	19	19
num_ret	27277933040	3040	6080	6080	2442335	3040	3040	5856	5856	
num_rel	1130	1130	1130	1130	1130	1130	1130	1130	1130	1130
num_rel_ret	1128	23	23	33	33	1120	193	193	255	255
map	0.0153	0.0072	0.0072	0.0081	0.0081	0.0976	0.0637	0.0637	0.0709	0.0709
P@5	0.0421	**0.0632**	**0.0632**	**0.0632**	**0.0632**	0.3263	0.3053	0.3053	0.3053	0.3053
P@10	0.0316	**0.0474**	**0.0474**	**0.0368**	**0.0368**	0.3	0.2895	0.2895	0.3	0.3
P@15	0.0246	**0.0386**	**0.0386**	**0.0351**	**0.0351**	0.2877	0.2702	0.2702	0.2772	0.2772
P@20	0.0263	**0.0342**	**0.0342**	**0.0368**	**0.0368**	0.2447	0.2395	0.2395	0.2447	0.2447
P@30	0.0246	**0.0263**	**0.0263**	**0.0316**	**0.0316**	0.214	0.2018	0.2018	0.2123	0.2123

Table 2. Queries with low number of relevant documents

runid	IN BM25 Low	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*	DOC BM25 Low	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*
num_q	46	46	46	46	46	46	46	46	46	46
num_ret	64606057200	7200	14400	14400	4015889	7185	7185	13953	13952	
num_rel	246	268	246	246	246	246	246	246	246	246
num_rel_ret	246	3	2	7	7	238	94	94	112	111
map	0.0027	0.0001	0.0001	0.0016	0.0016	0.0745	0.0792	0.0792	0.0805	0.0694
P@5	0	0	0	0	0	0.08	**0.0844**	**0.0844**	**0.0844**	0.08
P@10	0.0022	0	0	0.0022	0.0022	0.0689	**0.0756**	**0.0756**	**0.0756**	**0.0733**
P@15	0.0015	0.0015	0	0.0015	0.0015	0.0593	**0.0696**	**0.0696**	**0.0681**	**0.0667**
P@20	0.0022	0.0011	0	0.0022	0.0022	0.0544	**0.0633**	**0.0633**	**0.0611**	**0.06**
P@30	0.0015	0.0007	0	0.0015	0.0015	0.0489	0.0489	0.0489	**0.0519**	**0.0511**

Results The iSearch collection consists of queries with a high and a low number of relevant documents. It turns out the number of relevant documents for a query has an effect on the results. To document this, we provide figures for queries with a high and a low number of relevant documents as well as for all queries. The evaluation results for information need and document-based polyrepresentation are shown in Table 1 for the queries where positive relevance judgments for 20 or more documents are available (we call this *High*). The Table 2 holds the results for the queries where positive relevance assessment for less than 20 documents were available, we refer to this as *Low*. The results for all the queries combined are presented in Table 3, we refer to that as *All*. In the left half the tables the results for information need based polrepresentation are given and the right half holds the results for document and context-based polyrepresentation. In Table1 for information need based polyrepresentation, the proposed method performs better than the BM25 baseline (IN BM25) at *prec*@*n*, *prec*@10, *prec*@20 and so on, for both *eF* and *SD*, for $k = 5$ and $k = 10$. For the document-based polyrepresentation part the performance of *eF* and *SD* at *prec*@5, *prec*@10, *prec*@20 for both $k = 5$ and $k = 10$ is poorer than the baseline (DOC BM25). In Table 2 for the *Low* queries for $k = 5$ and $k = 10$ the values for all *prec*@*n* are very poor compared the baseline for information need polyrepresentation. In the document-based polyrepresentation part the *eF* and *SD* for both $k = 5$ and $k = 10$ show a performance improvement over the baseline at all *prec*@*n*. In Table 3 for the information need part there is no improvement, for the document-based

polyrepresentation part we can see a slight improvement at *prec*@5, *prec*@10, *prec*@15 and *prec*@20 but at *prec*@30 the performance is bit poorer.

Discussion The results show that our approach is able to improve the effectiveness, but there are interesting differences when it comes to queries with a low and a high number of relevant documents. For queries with high numbers of relevant documents, polyrepresentative clustering based on documents and bibliographic context does not perform well, whereas information need polyrepresentation reacts well on clustering. It seems that with polyrepresentation of information needs relevant documents are not necessarily found in the total cognitive overlap and are unearthed by means of clustering. An explanation may be that some combinations of information needs representations are not beneficial [19] and clustering is a means to mitigate this. We see a different picture for queries with only few relevant documents. Here, document-based polyrepresentative clustering outperforms the polyrepresentation baseline, but information need based clustering does not. For this reason the approach delivers a rather mixed result when we consider all queries. We also observe that clustering has a negative effect on mean average precision (map) and recall, which is due to the fact that we only select the top k documents from each cluster and drop the rest.

Table 3. All the queries

runid	IN BM25 All	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*	DOC BM25 All	eF *k=5*	SD *k=5*	eF *k=10*	SD *k=10*
num_q	65	65	65	65	65	65	65	65	65	65
num_ret	918841	10240	10240	20480	20480	5708875	10225	10225	19809	19808
num_rel	1376	1376	1376	1376	1376	1376	1376	1376	1376	1376
num_rel_ret	1376	25	25	40	40	1317	287	287	367	366
map	0.007	0.0022	0.0022	0.0035	0.0035	0.0816	0.0746	0.0746	0.0776	0.0698
P@5	0.0187	0.0187	0.0187	0.0187	0.0187	0.1469	**0.15**	**0.15**	**0.15**	0.1469
P@10	0.0125	0.0141	0.0141	0.0125	0.0125	0.1375	**0.1391**	**0.1391**	**0.1422**	**0.1406**
P@15	0.0104	0.0115	0.0115	0.0115	0.0115	0.124	**0.1292**	**0.1292**	**0.1302**	**0.1292**
P@20	0.0102	0.0102	0.0102	**0.0125**	**0.0125**	0.1117	**0.1156**	**0.1156**	**0.1156**	**0.1148**
P@30	0.0094	0.0078	0.0078	**0.0104**	**0.0104**	0.1	0.0943	0.0943	0.0995	0.099

4 Conclusion

In this study we have evaluated the suitability of combining principle of polyrepresentation with document clustering. The idea is that instead of a ranked list we present the user with clustering reflecting the relevance of documents w.r.t. different representations. We report some insights on the performance of adding bibliographic information as the representation as well as looking at information need polyrepresentation. Apparently the aspect of presenting the results to the user when various representations of information need as well as information object come into play is a complex task. The initial exploration show that

presenting information to the user in illustrated way has its potentials and drawbacks. We intend to test the approach for variable size k and the comparison of the cluster ranking strategies, i.e. arithmetic mean, geometric mean etc, in the future.

References

1. Ingwersen, P., Järvelin, K.: The turn: integration of information seeking and retrieval in context. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005)
2. Kelly, D., Fu, X.: Eliciting better information need descriptions from users of information search systems. Information Processing & Management **43**(1) (2007) 30–46
3. Skov, M., Larsen, B., Ingwersen, P.: Inter and intra-document contexts applied in polyrepresentation for best match IR. Information Processing & Management **44**(5) (2008) 1673–1683
4. Larsen, B., Ingwersen, P., Kekäläinen, J.: The polyrepresentation continuum in IR. In: Proceedings IIiX 2006, New York, NY, USA, ACM (2006) 88–96
5. Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., van Rijsbergen, K.: Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework. In: Proceedings IIiX 2010, ACM (2010) 115–124
6. Frommholz, I., Abbasi, M.K.: On clustering and polyrepresentation. In: Proceedings ECIR2014. To appear.
7. Borgman, C., Furner, J.: Scholarly communication and bibliometrics. Annual Review of Information Science and Technology **36** (2002) 3–72
8. Garfield, E.: Citation analysis as a tool in journal evaluation. Science (New York, N.Y.) **178**(4060) (November 1972) 471–9
9. Chen, C., Ibekwe-SanJuan, F., and Jianhua Hou: The structure and dynamics of cocitation clusters: A multipleperspective cocitation analysis. JASIST **61**(7) (2010) 1386–1409
10. Mayr, P., Mutschke, P.: Bibliometric-enhanced retrieval models for big scholarly information systems. IEEE International Conference on Big Data (2013)
11. White, H.: Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science. JASIST **58**(4) (2007) 583–605
12. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized blockmodeling. (2005)
13. Mutschke, P., Mayr, P., Schaer, P., Sure, Y.: Science models as value-added services for scholarly information systems. Scientometrics **89**(1) (June 2011) 349–364
14. Larsen, B., Ingwersen, P.: The Boomerang Effect : Retrieving Scientific Documents via the Network of References and Citations. (2002) 2–3
15. Fuhr, N., Lechtenfeld, M., Stein, B., Gollub, T.: The Optimum Clustering Framework: Implementing the Cluster Hypothesis. Information Retrieval **14** (2011)
16. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a Test Collection for the Evaluation of Integrated Search. In: Proceedings ECIR 2010. (2010) 627–630
17. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings SIGIR Workshop on Open Source Information Retrieval (OSIR 2006). (2006)
18. Wu, S.: Data Fusion in Information Retrieval. Volume 13. Springer (2012)
19. Lioma, C., Larsen, B., Ingwersen, P.: Preliminary experiments using subjective logic for the polyrepresentation of information needs. In: Proceedings IIiX 2012, ACM (2012) 174–183