Polyrepresentative Clustering: A Study of Simulated User Strategies and Representations

Muhammad Kamran Abbasi and Ingo Frommholz

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

Abstract. The principle of polyrepresentation and document clustering are two established methods for Interactive Information Retrieval, which have been used separately so far. In this paper we discuss a cluster based polyrepresentation approach for information need and document based representations. In our work we simulate and evaluate two possible cluster browsing strategies a user could apply to explore the polyrepresentative clusters. In our evaluation we apply information need and bibliographic features on the iSearch collection. Our results suggest that polyrepresentative cluster browsing may be more effective than exploring a ranked list.

1 Introduction

Interactive Information Retrieval (IIR) systems are supposed to support users to satisfy their information need beyond typing in queries. Unlike traditional rank based retrieval, interactive systems improve the user's search experience by providing extended means for user interactions in the overall search process. To this end, *polyrepresentation* has been identified as an important principle in IR [1]. The principle suggests that if multiple cognitively different (coming from different users) representations (i.e., reviews, ratings etc.) and functionally (coming from the same user for different purpose) different representations (i.e., title, abstract, references etc.) point to an information object then it is likely to be relevant to the user's information need. This situation 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*, the set of documents relevant w.r.t. all representations. $\mathcal{R}_0$ the set of documents not relevant to any representation at all. According to the principle of polyrepresentation this set is supposed to hold the most relevant documents as evaluated in [2]. This notion is shown in Figure 1. The principle of polyrepresentation has been evaluated so far in ad hoc retrieval and is used to create a ranked list. The actual challenge lies in modelling and reflecting user preferences in the context of

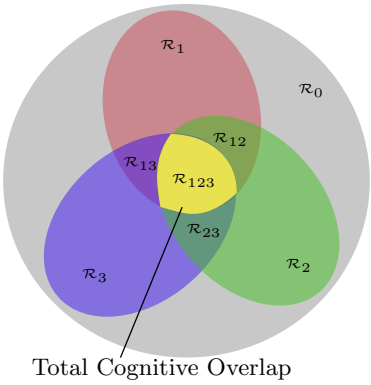

Fig. 1. Polyrepresentation overlaps

the principle of polyrepresentation. 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 these representations – for instance, is a user interested in the title but not references? A binary overlap computation found in most of the literature is not sufficient to cover such a varying importance of the representations. One approach to mitigate this is to let the user decide about their preference during the search session. To this end we propose to combine document clustering and polyrepresentation. From this we can derive a "polyrepresentation cluster hypothesis"– documents relevant to the same representations should appear in the same cluster [3]. Thus, we argue that instead of a ranked list, users may be presented with the clusters which represent the overlap shown in Fig. 1. By browsing the clusters and thus determining the sequence of visited clusters, the user got stronger means at hand to explore the results according to their interests. In this study an attempt is made to explore the potential and limitations of cluster-based polyrepresentation as an alternative to a result list organised in a single sequential list.

In Section 2 the cluster-based polyrepresentation is highlighted and in Section 3 our evaluation. A discussion and conclusion in given in Section 5.

2 Polyrepresentation and Document Clustering

The document clustering techniques in IR are proven effective for enhancing the overall search process [4]. They drive their justification from the well-known cluster hypothesis [5]. It is argued in the literature that clustering helps users in interactive information retrieval when it is difficult for them to specify information needs [6]. In contrast to the traditional heuristic clustering approaches, an Optimum Clustering Framework (OCF) has been proposed [6], with a sound theoretical justification for probabilistic document clustering. The framework utilizes so-called *query sets*, by following the cluster hypothesis in a reversed order: documents relevant to same queries should appear in the same cluster.

In a nutshell, OCF-based clustering represents each document d as a vector $\tau(d) = (P(R|d, q_1), \ldots, P(R|d, q_n))$ of the probability of relevance of the document with respect to each query in the query set. Traditional clustering algorithms can then be applied to this representation. This notion goes along with polyrepresentation based document clustering as described in [3], which states that the polyrepresentative overlaps as described in Figure 1 could be generated with the help of document clustering. In order to implement document-based polyrepresentation clustering, we employ the OCF-based notion of query sets. Thus for our polyrepresentation based clustering approach we intend to discover the possible clusters for the polyrepresentative sets $\mathcal{R}$ by estimating the degree of overlap, in this case the probability of relevance of each representation to the overlap. The τ vectors for information need based representations $r_i \in REP_{in}$ for document d become $\tau_{in}(d) = (P(R|d, r_1), \ldots, P(d, r_n))$. Similarly, the vector for document-based representation $rd_i \in REP_{doc}$ becomes $\tau_{doc}(d) = (P(R|rd_1, q), \ldots, P(R|rd_m, q))$. In previous work we evaluated the effectiveness of polyrepresentative clustering using τ_{in} and τ_{doc} separately [7]. One of the contributions of this work is to evaluate the concatenation of both vectors to a vector $\tau_{(in\ doc)} \in \mathbb{R}^{n+m}$ with

$$\tau_{(in\ doc)}(d) = (P(R|d, r_1), \ldots, P(d, r_n), P(R|rd_1, q), \ldots, P(R|rd_m, q)) .$$

This way we combine document and information need polyrepresentation, which to our knowledge has not been tried before, the motivation to combining REP_{in} and REP_{doc} was to explore whether such a combination contribute in improving retrieval. The τ vectors can then be used to cluster the documents with a suitable clustering function. Furthermore the single probabilities of relevance can be combined to create a within-cluster ranking that the user can explore. The evaluation of a cluster-based polyrepresentation approach poses many challenges, among them are finding the total cognitive overlap cluster, cluster order and the number of picked document from each cluster.

In [3] it is identified that cluster ranking methods are helpful in identifying the candidate cluster for the total cognitive overlap, further in [7] the initial evaluation of the polyrepresentative clustering approach is presented for information need based polyrepresentation and document based polyrepresentation. In this work we explore the effects of cluster based polyrepresentation approach on the combination of both the information need and document based representations, as expressed in the document vector $\tau_{(in\ doc)}$ and a further combination method described in the next section. This should give us an idea if a richer representation is beneficial for our approach. In Section 3.3 we present strategies to simulate the user behaviour, i.e. some naive and basic browsing strategies. Here the goal is to demonstrate that offering cluster-based interaction means in a polyrepresentative environment can indeed lead to a more effective search.

3 Evaluation

3.1 Collection

The goal of our evaluation is to demonstrate if a richer representation, including both information need and document polyrepresentation, leads to further improvements. We also look at applying different user-based exploration strategies. To conduct our experiments, the PF part of the iSearch [8] collection is used. This sub-collection contains full text physics articles. The collection comprises 65 search tasks, where each search task is expressed in five information need (IN) representations, i.e., Search Terms (st), Work Task (wt), Current Information Need (cn), Ideal Answer (ia) and Background Knowledge (bk). These five representations build the REP_{in} part of the experiments. For document based polyrepresentation; Title (ti), Abstract (ab), Body Text (bt), References (re) representations were extracted from full text articles. An additional context representation has been built based on the citation data available in the collection as described in [7]. All these representations constitute the REP_{doc} part of the experiment collection. We report the results for the concatenation of the REP_{in} and REP_{doc} (referred to as REP_{conc}). Furthermore, as previous experiments in polyrepresentation suggest not all representations may be equally effective, we look at pairwise combinations of individual representations comprised in REP_{in} and REP_{doc}, respectively. We refer to this as REP_{comb}. For example, (ti ab) denotes the combination of title and abstract, (ct re) means context combined with references. In order to estimate the $P(R|d, r_i)$ and $P(R|dr_i, q)$ in absence of actual relevance judgements for each representation, the BM25 based documents weights have been computed for every information need representation REP_{in_i} and each document representation REP_{doc_i}, respectively, using the Terrier IR platform [9]. These weights constitute the document vectors τ for clustering, as described in Section 2. To cluster the document vectors we used a standard *k-means* implementation in Matlab2011 with 'cityblock' distance while setting k to $2^{|REP|}$ (the motivation for computing $2^{|REP|}$ clusters was to match the number of possible overlaps as shown in Fig. 1 for $k = 3$). The k for $REP_{conc}$was set to 2^{10} and for REP_{comb} to 2^2 (as we only look at representation pairs. As a baseline for REP_{conc} we created a ranked list from the actual BM25 document weights for all 10 representations fused together using the CombSum fusion method [10]. Similarly, for REP_{comb} the BM25 weights for the two representations in the pair were combined together using CombSum to create the respective BM25 baseline. The same strategy was applied to create a within-cluster ranking of documents belonging to a cluster. In our experiments we used graded scale relevance for computing $NDCG@k$; for $P@k$ the 4-point graded scale was reduced to binary values, i.e., $rel = 0$ was mapped to non relevant and $rel > 0$ was mapped to relevant. We map n-tier relevance values to binary ones as one of the user strategies below assumes that a binary judgement is made.

[0] http://itlab.dbit.dk/~isearch/

3.2 Cluster Ranking

To simulate the user behaviour regarding the possible sequence in which clusters could be picked by the user or presented to the user, we ranked clusters using different criteria. The motivation of choice of such criteria was to use only information available in the cluster without relying on some external cluster quality measure. The two of such criteria were *arithmetic mean* and *geometric mean* as described in [11]. The arithmetic mean of a cluster C was computed as: $arith(C) = \frac{1}{|C|}\sum_{d\in C}\sum_{i=1}^{n}\frac{Pr(R|d,r_i)}{n}$, while the geometric mean of a cluster C was computed over the summed scores of the documents in the cluster as $geom(C) = \left(\prod_{d\in C}\sum_{i=1}^{n}\frac{Pr(R|d,r_i)}{n}\right)^{\frac{1}{|C|}}$. Besides these the OCF based (*expected F-measure (eF)*)[6] was derived as follows. For a cluster C in the clustering $\mathcal{C}$ let $\sigma(C) = \frac{1}{|C|-1}\sum_{(d_l,d_m)\in C_i\times C_i}\tau(d_l)^T \times \tau(d_m)(l \neq m)$ if $|C| > 1$, and 0 otherwise. Then the *expected pairwise precision of C is defined as* $\pi(C) = |C|\sigma(C)$. Likewise, the *expected recall* is defined as $\rho(C) = |C_i|(|C_i| - 1)\sigma(C_i)$. Based on these the *expected F-measure* is computed as: $eF = \frac{2}{\frac{1}{\pi(C)}+\frac{1}{\rho(C)}}$. Besides these measures we used the *sparsity density*, which is computed over the document×representation matrix which constitutes the cluster. For example, if a cluster C holds $|C|$ documents having $|REP|$ representations then this makes a $|C| \times |REP|$ matrix M_c where $P(R|d, r_i)$ becomes an element of matrix M_c. Thus, the sparsity of the M_c becomes the number of non-zero values in the matrix i.e., $P(R|d, r_i) > 0$, which could be denoted as $|M_c > 0|$. We then divided it with the total number of elements in the cluster matrix $|M_c|$ to indicate the Sparsity Density as: $SD(C) = \frac{|M_c>0|}{|M_c|}$.

3.3 Cluster Browsing Strategies

In order to evaluate the polyrepresentation based document clustering approach we use a simulated user methodology [12]. The first strategy *strategy-1* is described in [7], where for each query q the user is assumed to look at top l documents from each cluster. The sequence of clusters the user is visiting and the l documents in each within-cluster ranked list examined by the user create an *artificial ranked list* made of the documents and their sequence the user is assumed to be visiting them when exploring the polyrepresentative clusters. To simulate this, the clusters need to be ranked on the basis of some cluster quality criteria. The first cluster to be presented to the user is supposed to represent the total cognitive overlap as it is assumed to contain documents relevant to all representations involved. To check if a cluster browsing strategy is more effective than browsing a single ranked list, the resulting artificial cluster based ranked list is re-ranked and is evaluated against the actual BM25 ranking (our baseline as discussed above).

The second simulated user strategy, *strategy-2*, is as follows: we apply cluster ranking to simulate the sequence of clusters the user is visiting. We also create a ranked list of documents in each cluster as described above. The first cluster will

be presented to the user; from its ranked list the user examines the first document and looks at the second document only if the previously visited document would be relevant (hence the binary relevance judgements described above). This procedure continues until the user comes across a non-relevant document in the cluster. When user encounters a non-relevant document, user moves on to the next cluster in the cluster rank. For each cluster this procedure is repeated – the user is assumed to examine the documents in the within-cluster ranking until a non-relevant document is reached and proceeds to the next cluster. In any case the first document of each cluster is added to the ranked list. Again we can create an artificial ranked list from the visited documents re-rank them and evaluate them against the BM25 baseline.

4 Results and Discussion

For *strategy-1* and REP_{conc} we compared the created cluster based ranked list against their BM25 baseline, in Table 1 and 2, the entries in bold show the average performance improvement. The entries marked with (*) are statistically significant based on two tailed paired sample t-test at 95% confidence intervals. The performance improvement for *strategy-1* regarding REP concatenated to some extent confirms that the multiple representations of functionally and cognitively different nature could be useful for the performance benefit. But we can also observe a rather negative effect on the overall performance when we concatenate IN and document representations – the results for REP_{conc} lie between the values for separate document and IN polyrepresentation that were reported in [7]. Given the lower overall results for IN based polyrepresentation, this could have been expected. However, it should be noted that for REP_{conc} we were able to beat the respective BM25 baseline significantly for NDCG@30 and P@30, although these values are still below the ones for individual document polyrepresentation.

Table 1. *strategy-1* concatenated REP_{in} and REP_{doc} representations $P@k$ bold values show improvement over baseline

$l = 5$	BM25	arithMean	eF	geomMean	SD
P@5	0.0769	0.0769	0.0769	0.0769	0.0769
P@10	0.0462	**0.0477**	**0.0477**	**0.0477**	**0.0477**
P@15	0.0359	**0.0390**	**0.039**	**0.0390**	**0.0390**
P@20	0.0323	**0.0354**	**0.0354**	**0.0354**	**0.0354**
P@30	0.0256	**0.0313***	**0.0313***	**0.0313***	**0.0313***

The evaluation results for *strategy-2*, described in Section 3.3 are given in Table 3 for both $P@k$ and $NDCG@k$. Performance of *strategy-2* remains better than the baseline on average, but the overall improvement is not statistically significant. The actual set back for our *strategy-2* of user interaction is the very strict assumption that the user moves to a different cluster after observing the first non-relevant document. But even if the top-ranked document is not relevant, those appearing at second and third position in the cluster document rank could

Table 2. *strategy-1* concatenated REP_{in} and REP_{Doc} representations $NDCG@k$ bold values show improvement over baseline

$l = 5$	BM25	arithMean	eF	geomMean	SD
NDCG@5	0.0362	0.0362	0.0362	0.0407	0.0362
NDCG@10	0.0399	**0.0407**	**0.0407**	**0.0407**	**0.0407**
NDCG@15	0.0433	**0.0453**	**0.0453**	**0.0453**	**0.0453**
NDCG@20	0.0474	**0.0500**	**0.0500**	**0.0500**	**0.0500**
NDCG@30	0.0507	**0.0591***	**0.0591***	**0.0591***	**0.0591** *

be relevant, which by this assumption are ignored. However, the results suggest improvements are possible with a more refined strategy.

Table 3. *strategy-2* concatenated REP_{in} and REP_{Doc} representation $P@k$ and $NDCG@k$

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0362	0.0399	0.0433	0.0474	0.0507
strategy-2	**0.0375**	**0.0430**	**0.0499**	**0.0539**	**0.0566**
	p@5	p@10	p@15	p@20	p@30
BM25	0.0769	0.0462	0.0359	0.0323	0.0256
strategy-2	0.0677	**0.0477**	**0.0390**	**0.0354**	**0.0282**

In Table 4 the REP_{comb} results for *strategy-1* are given for selected document representation pairs. Overall we observe improvements over the results for all document representations reported in [7], which confirms previous work that certain representations do not contribute positively. However, further improvements seem to be possible when applying a cluster-based strategy (*strategy-1* in this case).

Table 4. *strategy-1* Representation combinations for REP_{doc} based polyrepresentation

REP_{doc}		BM25	arithMean	geomMean	SD
(ti ab)	P@5	0.154	0.154	0.154	0.154
	P@10	0.111	**0.118**	**0.118**	**0.118**
	NDCG@5	0.077	0.077	0.077	0.077
	NDCG@10	0.093	**0.098**	**0.098**	**0.098**
(ab ct)	P@5	0.132	0.132	0.132	0.132
	P@10	0.098	**0.102**	**0.102**	**0.102**
	NDCG@5	0.062	0.062	0.062	0.062
	NDCG@10	0.074	**0.079**	**0.079**	**0.079**
(ct re)	P@5	0.095	0.095	0.095	0.095
	P@10	0.072	**0.080**	**0.080**	**0.080**
	NDCG@5	0.053	0.053	0.053	0.053
	NDCG@10	0.060	**0.065**	**0.065**	**0.065**

5 Conclusion

In this paper, a cluster-based polyrepresentative approach for IN and document based representations has been explored along with the cluster browsing

strategies for simulating the user interaction in context. The concatenations and combinations of various representations were compared against respective BM25 document rankings. The evaluation results confirm findings in previous studies that cluster-based polyrepresentation has potential benefits in interactive IR. However, a concatenation of IN and document based polyrepresentation does not seem to be useful as results are somewhat between those for IN and document polyrepresentation alone. Moreover the findings reveal that different combinations of fewer individual representations do improve performance and a cluster-based approach seems promising here as well. The cluster browsing *strategy-2* appears useful, however need to be adapted further as it turns out to be too strict. In our future work we will refine and apply different cluster browsing strategies, based on *strategy-1* and *strategy-2* discussed here. We will also look at a different strategy to combine IN and document based polyrepresentation, where each IN representation is matched against each document representation.

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. 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
3. Frommholz, I., Abbasi, M.: On clustering and polyrepresentation. In: Proceedings ECIR2014. (2014) 618–623
4. Tombros, A., Villa, R., Van Rijsbergen, C.: The effectiveness of query-specific hierarchic clustering in information retrieval. Information Processing & Management **38**(4) (2002) 559–582
5. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)
6. Fuhr, N., Lechtenfeld, M., Stein, B., Gollub, T.: The Optimum Clustering Framework: Implementing the Cluster Hypothesis. Information Retrieval **15**(2) (2011) 93–115
7. Abbasi, M.K., Frommholz, I.: Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval. (2014) 21–28
8. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In: Advances in Information Retrieval. Springer (2010) 627–630
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)
10. Wu, S.: Data Fusion in Information Retrieval. Springer, Heidelberg (2012)
11. Kurland, O., Raiber, F., Shtok, A.: Query-Performance Prediction and Cluster Ranking: Two Sides of the Same Coin. In: Proceedings of the 21st ACM international Conference on Information and Knowledge Management - CIKM '12. (2012) 2459–2462
12. Azzopardi, L.: The economics in interactive information retrieval. In: Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), New York, New York, USA, ACM Press (2011) 15–24