Scientometrics (2015) 102:2301–2322
DOI 10.1007/s11192-014-1478-1

Cluster-based polyrepresentation as science modelling approach for information retrieval

Muhammad Kamran Abbasi · Ingo Frommholz

Received: 18 October 2014 / Published online: 25 November 2014

Abstract The increasing number of publications make searching and accessing the produced literature a challenging task. A recent development in bibliographic databases is to use advanced information retrieval techniques in combination with bibliographic means like citations. In this work we will present an approach that combines a cognitive information retrieval framework based on the principle of polyrepresentation with document clustering to enable the user to explore a collection more interactively than by just examining a ranked result list. Our approach uses information need representations as well as different document representations including citations. To evaluate our ideas we employ a simulated user strategy utilising a cluster ranking approach. We report on the possible effectiveness of our approach and on several strategies how users can achieve a higher search effectiveness through cluster browsing. Our results confirm that our proposed polyrepresentative cluster browsing strategy can in principle significantly improve the search effectiveness. However, further evaluations including a more refined user simulation are needed.

Keywords Information retrieval · Polyrepresentation · Document clustering · Bibliometrics · Simulated user

Introduction

The increasing number of publications in almost all domains of knowledge makes searching and accessing information about the produced literature a challenging task. Online bibliographic databases help to identify the information about authorship/co-authorship, collaboration, citation as well as the impact of research produced in any

M. K. Abbasi (✉) · I. Frommholz
Institute for Research in Applicable Computing, University of Bedfordshire, Park Square, Luton, LU1 3JU, UK
e-mail: abbasikamran@gmail.com

I. Frommholz
e-mail: ingo.frommholz@beds.ac.uk

domain. The meta information of this kind provides a rich source of evidence that could be used to improve searching and accessing relevant information from bibliographic databases. However, the specific nature of this information needs proper treatment to be useful for digging down the connections between information objects, their relevance to each other and their retrieval. Despite this rich source of potential evidence for the relevance of a document to a user's information needs, models exploiting this resource for information retrieval (IR) in bibliographic databases are still rare. In particular interactive information retrieval (IIR) is supposed to support the user beyond just typing in queries. In this paper, we will describe how some methods from IIR, namely polyrepresentation and clustering, can be used in combination with bibliometrics.

In IIR, the principle of *polyrepresentation* is an important cognitive approach as it suggests to use the various information representations available for an information object in context to enhance the retrieval quality (Ingwersen 1996; Ingwersen and Järvelin 2005). The representations could be cognitively different, which means they are created or coming from a different actor in a different context, or they could be functionally different, which means they are coming from the same actor, but serve a different purpose. An example for cognitively different representations of information objects are the document content, coming from an author in a certain context, and reviews of a document, coming from different users that share their perspective of the document under consideration with their particular context or task as background. Citations are another example of cognitively different representations. If a document d_1 cites a document d_2, a re-interpretation of (some) content of d_2 is provided in d_1. Also d_2 is related to d_1 and possibly displays a different view about content discussed in d_1. We can argue that (parts of) d_1 are a different interpretation of (parts of) d_2 and vice versa. We will apply this kind of cognitively different representation in our polyrepresentative clustering approach discussed later. A functionally different representation example would be the title and the abstract of a document, which serve a different purpose, but are from the same actor (the author). The examples so far comprise polyrepresentation of information objects. The cognitive context of actors does not only manifest in information objects like documents, but can also be expressed by describing or representing the user's information need in different ways. Here the query is the most obvious representation, but others are possible, like the working context or task, a textual information need description or a description of the user's background knowledge.

In our work, we will make use of these two kinds of polyrepresentation (information objects/documents and information needs). Our idea is to combine polyrepresentation with document clustering, as it was suggested by Frommholz and Abbasi (2014), to allow for an interactive search strategy. We regard the cited articles as a new representation of a document under consideration.

The remainder of this article is as follows. In the next section we will discuss some related work in the areas of polyrepresentation, clustering and IR in bibliometrics/scientometrics. Afterwards we will discuss our main approach to combine a probabilistic clustering framework and polyrepresentation in order to incorporate bibliometric features. We will present the underlying user model, experiments made with the iSearch collection and discuss their results before we finally conclude.

Related work

In this section we briefly discuss related work in the areas of clustering as well as bibliometrics. An in-depth discussion of polyrepresentation will be provided in

"Polyrepresentative cluster browsing" section when we introduce our idea of a polyrepresentative cluster browsing strategy.

Clustering

The document clustering approaches in IR have their standing besides retrieval based on ranked lists. Since the formulation of the *cluster hypothesis* (van Rijsbergen 1979) many researchers focused on evaluating the document clustering approaches for IR in a diverse range of applications. A scatter/gather based approach for browsing the large collections has been proposed in Cutting et al. (1992) based on the table of content metaphor. The evaluation of scatter/gather in the context of the cluster hypothesis is given in Hearst and Pedersen (1996) where the authors compared the scatter/gather approach with similarity search for IIR and report significant improvements. An online clustering version of scatter/gather is presented in Ke et al. (2009). Document clustering approaches for IIR are further evaluated in Leuski (2001); in this study the authors propose to partition the result list into clusters and present the user a so-called *clustered list*. The clusters are presented to the user neither as a document list nor as textual description, but only a representative document from the cluster is put in the cluster list, with the intention that it serves as a guiding indicator for the user to decide the relevance of the cluster to the information need.

In contrast to collection-based clustering, where the whole collection is clustered, query-based clustering usually post-processes a result list that is given as a response to a query. Our approach can be classified into this category. Query-based hierarchical document clustering approaches are discussed in Tombros et al. (2002). Here the authors compared a query-based clustering of a retrieved ranked list with the clustering of the whole collection and report significant performance improvements of query-specific clustering over the static collection clustering. Ji and Xu (2006) suggest to use the user's prior knowledge to enhance the clustering performance, reporting a performance improvement over traditional clustering algorithms (k-means, normalize cut, and transductive SVM). In Na et al. (2007) an adaptive approach for the document clustering for query-based similarity is proposed, the authors evaluated the similarity measures from language modelling, and report the performance improvement over k-means. A probabilistic approach for full text document clustering is proposed in Goldszmidt and Sahami (1998) where the authors compute the probabilistic overlap between documents as a similarity measure and suggest approaches to estimate probabilities from cosine similarity scores. Query-specific clustering cannot only be used for creating clusters, but clusters can also be ranked and, based on the cluster ranking, a new ranking of the documents can be created [see for instance, Raiber and Kurland (2013)]. To evaluate our cluster-based approach we will apply query-specific clustering and re-ranking to simulate the user's search behaviour to make the resulting ranked list comparable to a baseline ranking.

While a lot of work has been performed in document clustering, still the proposed approaches have been rather heuristic and a unifying framework was missing. This has led to the formulation of the Optimum Clustering Framework (OCF) that is based on the probability of relevance of documents with respect to a given query set (Fuhr et al. 2011). Standard clustering approaches can be regarded as a special case of the OCF where each term in the collection is a query in the query set. We will discuss a polyrepresentative clustering approach based on the OCF in more detail in "Polyrepresentative cluster browsing" section. Fuhr et al. (2011) also provide a further overview of document clustering methods in the light of the OCF.

Bibliometrics/scientometrics and information retrieval

Structure based mapping and modeling techniques of scholarly activities based on statistical methods are known as science models and are used to improve the retrieval quality in scholarly IR (Mutschke et al. 2011). Besides this IR approaches in their own right are well researched, tested and applied on a diverse range of situations. Thus the combination of the approaches from IR with bibliometrics/scientometrics may lead to promising results in both domains (Mayr and Mutschke 2013) and our work contributes to this body of work by utilising citations as document representations. Some limitations of the IR techniques for IR systems i.e. the vagueness of the query terms, indexing and retrieval and ranking of the information object, are discussed in Mayr et al. (2008); these augmentations are termed as so called value-added services for scholarly information systems. The integration of science models, i.e. co-term relevance, bradfordizing and co-authorship models of re-ranking with the IR systems are presented in Mutschke et al. (2011).

In general the focus has been on the evaluation of the science models with the measures known from IR to evaluate the effects of ranking and re-ranking based on the core journal centrality (bradfordizing), author centrality, and the effects of query expansion with the co-words extracted from the documents of the initial query terms. Chen et al. (2010) present the perspective on co-citation analysis, where the authors cited together in a relevant domain are taken as key features and the smart cluster labelling mechanisms based on these features are elaborated. A framework for recommending terms for digital libraries and information systems is presented in Ritchie et al. (2006) and its application for reducing the term vagueness is discussed in Mayr et al. (2008) along with the re-ranking based on bradfordizing and co-author network analysis. A term suggestion approach based on the principle of polyrepresentation is presented in Schaer et al. (2012). This approach extends the term suggestion with the author names, and reports an increase in retrieval performance. A term recommendation and an interactive query expansion approach for digital libraries is highlighted in Lüke et al. (2013). A term boosting method for scientific book record retrieval based on meta data is presented in Larsen et al. (2012). In most of the scientometric studies the bibliometric meta characteristics of the scientific publications are taken into account but the lexical connections remain untouched (Glenisson et al. 2005b). The combination of bibliometric information and full-text in the scientometrics domain is presented in Glenisson et al. (2005a, b). Document clustering techniques were explored and the authors emphasized the use of hybrid methodologies, i.e. data mining and scientometrics to map the field of science. An important study combining the IR and the bibliometrics worth mentioning is (Larsen 2002). In this study, the usage of references and citations is demonstrated for improving the retrieval performance for scientific papers.

Keeping this context in mind, our study aims at exploring the potential of polyrepresentation and document clustering as a science model mapping approach for scholarly IR in the science domain.

Polyrepresentative cluster browsing

In this section we will discuss how the principle of polyrepresentation can be applied to support a browsing strategy. After further introducing polyrepresentation, we will describe how polyrepresentation and clustering can be combined to support interactive retrieval through browsing.

Polyrepresentative browsing

We will discuss how the principle of polyrepresentation can be applied for an interactive search strategy where users browse the result set according to different representations they deem important. We will discuss the principle of polyrepresentation first before we present the search strategy that is underlying our further considerations.

Polyrepresentation

The principle of polyrepresentation could be applied in multiple situations. For illustration purposes we present an example of polyrepresentation of documents as discussed in Frommholz et al. (2010). A user, in need of a "good introduction to quantum mechanics" visits an online book store where multiple representations of an information object (book) exists (for instance, the abstract, reviews, ratings, the tags provided by users, perhaps the full text and the bibliographic metadata). These representations contain and potentially satisfy different parts of the user's information need in many ways. Abstract, title and full text, for example, may be used to determine that the document is about the required subject (quantum mechanics) while reviews and ratings may tell us if the book is a good introduction. In this situation according to the principle of polyrepresentation if an information object is relevant to more representations, it is likely to be relevant to the user's information need. This scenario is presented in Fig. 1, here the sets $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$ denote the documents that are relevant to the representations in question (e.g., $\mathcal{R}_1$ may represent documents whose content is relevant to the query, $\mathcal{R}_2$ may denote documents with relevant reviews, etc). In this scenario the set $\mathcal{R}_{12}$ is the intersection of $\mathcal{R}_1$ and $\mathcal{R}_2$, i.e. documents that are relevant w.r.t. both respective representations. If the document is relevant to all the three representations then it appears in $\mathcal{R}_{123}$, which in this case makes the so-called *total cognitive overlap*. According to the principle of polyrepresentation this set may have high precision, which is confirmed by several experiments (Kelly and Fu 2007; Larsen et al. 2006; Skov et al. 2008). The set $\mathcal{R}_0$ contains the documents not relevant to any of the given representations or in other words completely irrelevant documents.

While this simple example discusses three representations for illustration purposes, we can easily extend it to any number of representations. In a similar way we could illustrate the polyrepresentation of information needs and the combination of information need and information object polyrepresentation.

Polyrepresentative browsing strategy

Although the principle of polyrepresentation has been confirmed in the literature, its actual application in a retrieval system brings with it some open problems. While it is clear that the user should check the total cognitive overlap, as this is likely to contain relevant documents, it is not straightforward which set to present next to the user—this depends, for instance, on the user's actual preferences, which is often not known to the system. For example, the user may or may not be interested in reviews, if we recall our book store example. If the user is not interested in the reviews, then documents with a high probability of relevance for reviews but not for any other representation could be ignored. This strategy is elaborated in the user scenario discussed in "Evaluation" section.

Referring back to the scenario in Fig. 1, let us assume we found a way to create the different partitions $\mathcal{R}$ (we will see later how we can at least approximate these partitions).

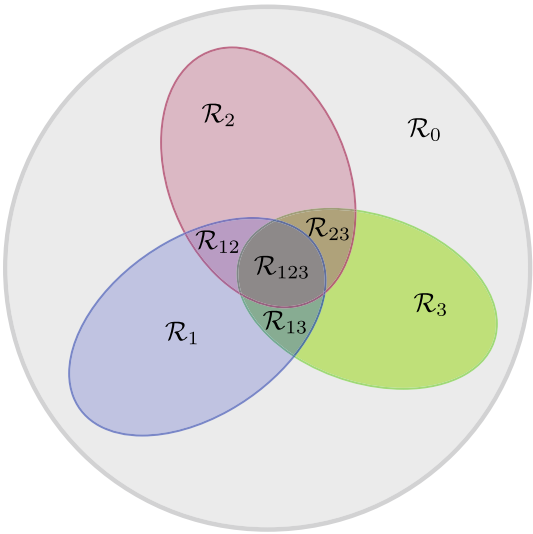

Fig. 1 Polyrepresentation-based relevance

As a search strategy, users may investigate the total cognitive overlap $\mathcal{R}_{123}$ first as the principle of polyrepresentation suggests. If the users are not interested in representation $\mathcal{R}_3$ but in the other representations they may now proceed to $\mathcal{R}_{12}$ and then later explore $\mathcal{R}_1$ and $\mathcal{R}_2$. This strategy imposes a weak ranking of representations provided by the user, in our case $\mathcal{R}_{123}$–$\mathcal{R}_{13}$–$(\mathcal{R}_1|\mathcal{R}_2)$. We may further assume the user does not investigate a whole partition, but only some top l documents in the respective partition. One of our claims is that such a polyrepresentative browsing strategy is more effective than exploring one single possibly polyrepresentative ranked list of documents.

Polyrepresentative clustering

The above browsing strategy assumed that we can somehow create our partitions $\mathcal{R}$ and present them to the user for exploration. The question that immediately arises is how this can be achieved. From the consideration above it becomes clear that polyrepresentation creates a partitioning of the document set based on representations. Furthermore, each document is contained in one and only one of the sets induced by polyrepresentation. If we assume each document can only be part of exactly one cluster, document clustering creates a similar partitioning of the document space. Naturally, we can ask if it is possible to create a polyrepresentation-induced partitioning by means of clustering where the clusters match the partitions $\mathcal{R}$.

As mentioned before, the OCF proposed by Fuhr et al. (2011) appears to provide a sound theoretical justification for document clustering in IR. The OCF is based on the well known cluster hypothesis (van Rijsbergen 1979). The OCF uses the notion of *query sets* by reversing the cluster hypothesis i.e. the documents relevant to the same queries in the query set appear in the same clusters. We present this idea for polyrepresentation in the form of a *polyrepresentation cluster hypothesis*: "documents relevant to the same representations should appear in the same cluster".

The OCF acts upon the probability of relevance $\Pr(R|d,q)$ of document d with respect to query $q \in \mathcal{Q}$ in the query set. Hence, each document d in a document set D is represented by a vector $\boldsymbol{\tau}$ as

$$\boldsymbol{\tau}(d) = \begin{pmatrix} \Pr(R|d,q_1) \\ \vdots \\ \Pr(R|d,q_n) \end{pmatrix} \tag{1}$$

where n is the number of queries in the query set $\mathcal{Q}$. Such document vectors are then clustered using any clustering function as per overall set up.

In order to use OCF with polyrepresentation we need to differentiate between the polyrepresentation of information needs and polyrepresentation of documents. In order to apply clustering to information need polyrepresentation let REP_{in} be the set of 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

$$\boldsymbol{\tau}_{in}(d) = \begin{pmatrix} \Pr(R|d,r_1) \\ \vdots \\ \Pr(R|d,r_n) \end{pmatrix} \tag{2}$$

with $n = |REP_{in}|$. $Pr(R|d,r_i)$ is the probability of relevance of the document d with respect to an information need representation r_i.

When applying polyrepresentation of documents or information objects, REP_d consists of the different representations rd_i of a document d. Here we assume that the information need is represented by the query q alone. We therefore need to compute $\Pr(R|rd_i,q)$ and we get

$$\boldsymbol{\tau}_{io}(d) = \begin{pmatrix} \Pr(R|rd_1,q) \\ \vdots \\ \Pr(R|rd_n,q) \end{pmatrix} \tag{3}$$

with $n = |REP_d|$.

In this paper we will focus on polyrepresentation of information objects and of information needs separately. However, to cover the full cognitive context of the user it could be interesting to combine representations of information needs with information object representation. In this case clustering would operate on the Cartesian product $REP_d \times REP_{in}$ and the document vector would be

$$\boldsymbol{\tau}_{io\times in}(d) = \begin{pmatrix} \Pr(R|rd_1,r_1) \\ \vdots \\ \Pr(R|rd_n,r_m) \end{pmatrix} \tag{4}$$

with $(rd_i,r_j) \in REP_d \times REP_{in}$ and $n = |REP_d|$, $m = |REP_{in}|$.

Having set up polyrepresentative clustering in this section we will now describe our experiments and the evaluation methodology.

Evaluation

Collection

The iSearch[1] collection (Lykke et al. 2010) is used to carry out experiments, which comes with a document corpus that comprises upon three sub-collections: meta-data for library book records (BK), the full text PDF documents (PF) and the meta-data with abstracts (PN). Furthermore, search tasks with five information need representations for each search task as well as corresponding relevance assessments also come with the collection. Additionally, there are 3.7 million direct citation in iSearch for the PF (110899) and PN (197783) sub-collections, and 12.7 million extracted references.

The initial goal of our evaluation was to assess the potential of principle of polyrepresentation when combined with a document clustering approach, in particular OCF, for scholarly IR and how it can go along with the science models, in particular references and citations. We focused on the full text (PF) sub-collection of iSearch. We use information need based polyrepresentation and the document based polyrepresentation.

Document vector creation and clustering

We describe how the document vectors $\boldsymbol{\tau}_{in}$ and $\boldsymbol{\tau}_{io}$ were created by means of information need and document polyrepresentation. $\boldsymbol{\tau}_{in}$ and $\boldsymbol{\tau}_{io}$ were clustered using *k-means* clustering (MacQueen et al. 1967). In order to be able to match the representation sets $\mathcal{R}$ we set k to $2^{|REP|}$ to produce as many clusters as there are representation sets.

Information need polyrepresentation

For information need based polyrepresentation, the information need representations provided with the iSearch collection were use to establish the set

$$REP_{in} = \{\text{search term}, \text{work task}, \text{background knowledge}, \text{ideal answer}, \text{current information need description}\}.$$

The information need representations were used as a query set along with the document full text to compute the $\Pr(R|d, r_i)$ used in Eq. 2. The PF sub collection and the parsed collection were indexed with Terrier 3.5[2] (Ounis et al. 2006). We estimated $\Pr(R|d, r_i)$, respectively, with BM25 (Robertson et al. 1998) as it was done in Fuhr et al. (2011). The BM25 weights were normalized by dividing each document weight with the highest weight computed for that particular representation.

Document polyrepresentation

For the document based polyrepresentation full-text articles were parsed to extract different sections i.e. title, abstract, body and references. The reference representation was constructed by taking the 'References' section of a paper and regard this as a textual representation. A further representation was the document context established by all articles cited by the article under consideration. The context of an article was created by

[1] http://itlab.dbit.dk/isearch/.

[2] http://terrier.org/.

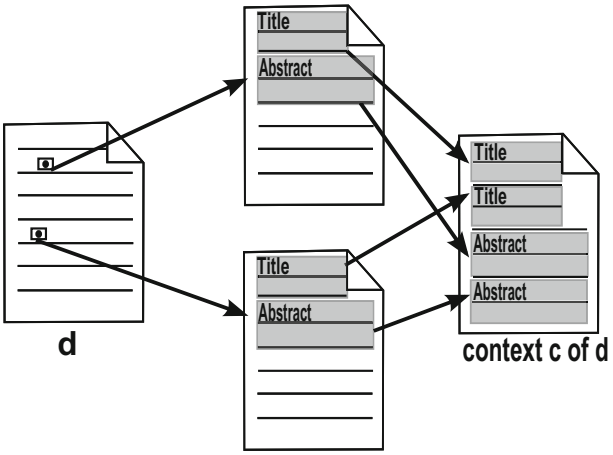

Fig. 2 Citation-based document context

merging the titles and abstracts of all cited articles as depicted in Fig. 2. This extraction was based on the direct citation data (for ≈110,900 full text articles in PF) provided with the collection. Clusters containing documents that are highly relevant w.r.t. references and the context provide the user the ability to explicitly choose the evidence for relevance coming from citations when visiting these clusters.

The representations make up the set REP_d to compute the $\Pr(R|rd_i, q)$ in Eq. 3 as

$$REP_d = \{\text{title}, \text{abstract}, \text{body}, \text{references}, \text{context}\}.$$

Again we estimated each $\Pr(R|rd_i, q)$ with BM25.

Simulated user methodology and cluster ranking

In order to evaluate our approach we will utilise a simulated user methodology, which was, for instance, applied in Azzopardi (2011) and other literature. We simulate the user behaviour in a very simple way as follows. The basic idea here is that for each information need (query), ranked clusters are presented to the user in a way that (s)he looks at top l documents in each cluster and then moves on to the next preferred cluster accordingly, where the user examines again the top l documents, and so on. In our experiments, we considered a static value for l as well as one based on the chosen cluster. In our evaluation, we determined the l in two ways, *fixed* l where l is static throughout the clusters and *variable* l where the l value is cluster-dependent. In our experiments, the value of the fixed l is set to 5 and 10 for all clusters. For the variable l we applied two strategies. In a first strategy we set $l = 10$ for the first cluster the user visits, $l = 8$ for the 2nd cluster. Generally, we apply a fixed sequence $10, 8, 6, 4, 2, 1, \ldots, 1$ for all $2^{|REP|}$ clusters we generate for setting l. We call this strategy *Variseq l*. Another strategy sets the l_i value for the $i+1$st visited cluster iteratively as $l_i = \lceil l_{i-1}/2 + 2 \rceil$ with $l_0 = 2^{|REP|}$. The top l_0 documents are selected from the first visited cluster, the top l_1 from the second visited cluster, and so on. The assumption is that users visit less documents the more clusters they have already looked at. We call this strategy *Varireps l*. It should be noted that these are some ad hoc search strategies that provide a simple simulation of the user's behaviour. More refined models should be based on user behaviour

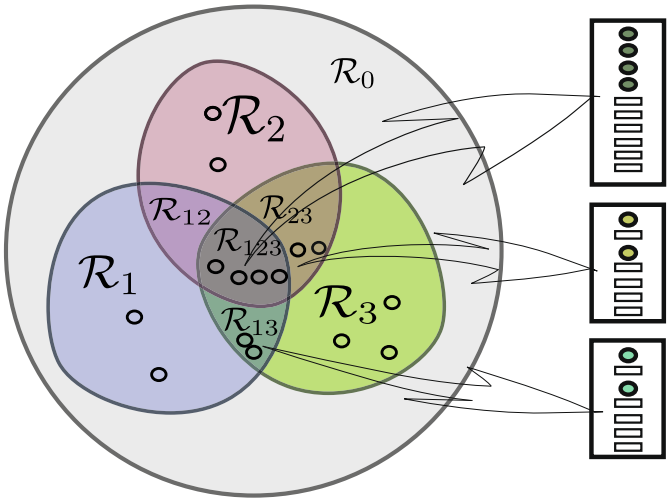

Fig. 3 Polyrepresentative cluster browsing. Assuming that each representation set $\mathcal{R}$ can be mapped to a cluster that contains a ranked list of documents, users explore the top l documents in the ranking

studies and will be subject to future work. The cluster browsing scenario is depicted in Fig. 3. In this figure, the cognitive overlaps are shown in a rank, where relevant documents are shown with circles and non relevant with rectangles.

It should be noted that by the way we simulate the user, a new ranking of documents is created based on the sequence of clusters examined and the within-cluster ranking. This way we can compare the interactive ranking approach against a baseline ranking, which may not be based on any clustering or even polyrepresentation, in a controlled environment utilising standard IR evaluation measures. All the documents the user looked at form a ranking according to the procedure given in Algorithm 1 for fixed l. For each query, the l documents from each clusters are combined together to create the ranking.

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

Algorithm 1: Cluster-based ranking for simulated user (fixed l)

The question that arises is how we determine which cluster the user chooses next. To this end, the computed clusters were ranked on the basis of different ranking measures, expected F-measure and sparsity-density. The *expected F-measure* is defined as

$$eF(D, \mathcal{Q}, C) = \frac{2}{\frac{1}{\pi(D,\mathcal{Q},C)} + \frac{1}{\rho(D,\mathcal{Q},C)}} \tag{5}$$

where π and ρ are the computed *expected precision* and *expected recall*, respectively, as defined in Fuhr et al. (2011) but on a per cluster basis. D is the set of documents, $\mathcal{Q}$ the query set (induced by the representations as discussed above) and C is the cluster under consideration. The other ranking measure used is *sparsity-density*, which is based on the matrix made of documents in a cluster and representations. If a cluster C contains $|C|$ documents and we are dealing with $|REP|$ representations, we can build a $|C| \times |REP|$ matrix M where each $\Pr(R|d, r_i)$ (or $\Pr(R|rd_i, q)$ in case we are using document polyrepresentation) is an element of. The idea behind the sparsity-density approach is to count the number of times we have got non-zero values in the matrix (i.e. $\Pr(R|d, r_i) > 0$ or $\Pr(R|rd_i, q) > 0$), denoted $|M_{>0}|$ and divide this by the number of elements in our matrix:

$$SD(C) = \frac{|M_{>0}|}{|M|}. \tag{6}$$

The *eF* measure is a cluster quality measure and the motivation to use *SD* is to find the total cognitive overlap (i.e., the cluster where all or many representations contribute with high scores)—if a cluster has many or all representations contributing then its SD score will be 1 whereas it approaches 0 when less or no representations contribute.

Experiments and results

In our experiments we evaluate our simulated user strategy that produces a ranking as described in Algorithm 1. We will investigate different strategies for l and for creating a cluster ranking. The created ranking is then compared to a BM25 baseline using polyrepresentation as follows. The BM25 values for all representations are computed to estimate $\Pr(R|rd_i, q)$ and $\Pr(R|d, r_i)$, respectively. We create our baseline ranking by combining the actual BM25 scores for all representations with CombSum (Fox and Shaw 1993). By using a polyrepresentative baseline we make sure that our clustering idea and the simulated user model is in the focus of our evaluation. We chose BM25 as this was used in other OCF-related experiments as an approximation of the probability of relevance (Fuhr et al. 2011).

We start our discussion with a general consideration of the potential of a cluster-based approach for polyrepresentation. To this end we generate an ideal scenario as presented next.

The ideal cluster ranking scenario

In order to validate the potential of the proposed method we designed an ideal cluster ranking scenario to see if any improvement can be achieved by means of the cluster ranking as proposed. We define the ideal cluster ranking as the ranking in which the clusters are ranked according to the absolute number of relevant documents in each cluster, which in this case could be equivalent to the ranking if human assessors are asked to rank the clusters which they consider relevant to some information need. This approach uses the relevance judgements provided with the iSearch collection. We extracted binary relevance judgements from the grades iSearch provides with a value >1 meaning relevance. Using the relevance judgements is of course not a realistic retrieval scenario. However, the

objective to use this kind of ranking is to test whether the proposed cluster ranking approach is worth exploring at all, with the hope that we can later devise cluster ranking approaches that come close to an ideal one. In this study, we report the precision at k (P@k) and NDCG at k (NDCG@k). Although in the literature precision at 10 is reported redundant in the presence of more complex measures, specially the average precision (AP) (Webber et al. 2008), we report precision at k, because the average precision is not reported and precision at k here is a supportive measure to NDCG at k, to present a clear picture. Also reporting precision at k makes our method comparable to the large body of work that utilises this measure.

The precision at k (P@k) and NDCG at k (NDCG@k) results of the ideal scenario are presented in Tables 1 and 2 for IN based polyrepresentation. For the ideal cluster ranking the dynamic part is the strategy to select documents from each cluster. We therefore analyse our fixed and variable strategies where l is set to 5, 10, and *Varireps l*, *Variseq l* as described in "Simulated user methodology and cluster ranking" section. The created ranks were evaluated using trec_eval, first for P@k and then for NDCG@k. The ranking results for each query were compared to the BM25 baseline for statistical significance. We compared the scores using paired sample Student's t test as described in Hull (1993), Smucker et al. (2007). For IN polyrepresentation we observe minor improvements, but no statistical significance can be reported here.

The Tables 3 and 4 show P@k and NDCG@k, respectively, for document based polyrepresentation. The ideal ranking results show significant improvements over the baseline everywhere with a slight tendency for the $l = 5$ strategy in case the user is interested in examining 5–10 documents in total. It seems that indeed relevant documents can be found within the first documents in relevant clusters, which speaks in favour of a cluster-based polyrepresentation search strategy, at least when it comes to document polyrepresentation. Relevant documents that would otherwise be lower in the ranking for instance with the BM25 strategy are now top-ranked documents in their respective cluster.

All in all the results for an ideal clustering are mixed but promising. For IN polyrepresentation we are able to produce slightly better results over a polyrepresentative baseline, but these are not statistically significant. IN polyrepresentation in general produces very low P@K and NDCG@K values, which needs to be further explored. Document polyrepresentation including bibliographic data on the other hand seems a very promising strategy as it produces significant improvements. It seems if users explore clusters rather than a ranked list they stand a chance to find relevant documents more effectively. The challenge is to point the user to the right clusters to explore. The results have motivated us to continue our exploration further in this direction.

Please note that in the tables discussed so far and some tables below, the P@5 and NDCG@5 values for our ideal cluster ranking are identical. This is due to the fact that in all our strategies, we take at least the first 5 documents from the best ranked cluster, so this does not come to a surprise.

Results of proposed method (all queries)

In this section we present the evaluation of the proposed method and discuss the results. The difference in these experiments is that we are not assuming an ideal cluster ranking based on existing relevance judgements to simulate the user's selection of clusters, but are applying automatic means to rank the clusters. In particular, the clusters are ranked using the eF and SD measures as described in "Simulated user methodology and cluster ranking" section and the ranked lists were created using Algorithm 1.

Table 1 Ideal cluster ranking: P@k for information need polyrepresentation

IN All	P@5	P@10	P@15	P@20	P@30
BM25	0.0187	0.0125	0.0104	0.0102	0.0094
Varireps l	0.0185	0.0123	**0.0133**	**0.0115**	**0.0113**
Variseq l	0.0185	0.0123	**0.0133**	**0.0108**	**0.0097**
$l = 5$	0.0185	**0.0138**	**0.0133**	**0.0123**	**0.0118**
$l = 10$	0.0185	0.0123	**0.0133**	**0.0146**	**0.0133**

Bold means improvement over the baseline

Table 2 Ideal cluster ranking: NDCG@k for information need polyrepresentation

IN All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.01195	0.0131
Varireps l	**0.0069**	**0.0119**	**0.0134**	**0.0148**	**0.0175**
Variseq l	**0.0069**	**0.0119**	**0.0134**	**0.0147**	**0.0167**
$l = 5$	**0.0069**	0.0075	0.0091	0.0097	0.0118
$l = 10$	**0.0069**	**0.0119**	**0.0124**	**0.0147**	**0.0167**

Bold means improvement over the baseline

Table 3 Ideal cluster ranking: P@k for document polyrepresentation

Doc All	P@5	P@10	P@15	P@20	P@30
BM25	0.1469	0.1375	0.1240	0.1117	0.1000
Varireps l	**0.2092**	**0.1677**	**0.1559**	**0.1354**	**0.1128**
Variseq l	**0.2092**	**0.1677**	**0.1539**	**0.1346**	**0.1108**
$l = 5$	**0.2092**	**0.1723**	**0.161**	**0.1392**	**0.1087**
$l = 10$	**0.2092**	**0.1677**	**0.1600**	**0.1346**	**0.1138**

Bold means statistical significance (with $p < 0.01$), and improvement over the baselin

Table 4 Ideal cluster ranking: NDCG@k for document polyrepresentation

Doc All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
Varireps l	**0.1411**	**0.1746**	**0.1997**	**0.2096**	**0.2274**
Variseq l	**0.1411**	**0.1746**	**0.1999**	**0.2104**	**0.2255**
$l = 5$	**0.1411**	**0.1809**	**0.2030**	**0.2159**	**0.2266**
$l = 10$	**0.1411**	**0.1746**	**0.2047**	**0.2118**	**0.2324**

Bold means statistical significance (with $p < 0.01$), and improvement over the baseline

Tables 5 and 6 show the precision and NDCG values, respectively, for different ranking positions k and information need polyrepresentation. When it comes to P@k we do not observe a difference between the eF and SD cluster ranking strategies. This slightly changes when we look at the more refined NDCG@k values, which reveal a slight preference for the SD technique. However, the improvements were not significant.

Table 5 P@k for information need polyrepresentation

IN All	P@5	P@10	P@15	P@20	P@30
BM25	0.0187	0.0125	0.0104	0.0102	0.0094
eF $l = 5$	0.0187	**0.0141**	**0.0115**	0.0102	0.0078
SD $l = 5$	0.0187	**0.0141**	**0.0115**	0.0102	0.0078
eF $l = 10$	0.0187	0.0125	**0.0115**	**0.0125**	**0.0104**
SD $l = 10$	0.0187	0.0125	**0.0115**	**0.0125**	**0.0104**
eF *Varireps l*	0.0187	0.0125	**0.0123**	**0.0115**	0.0087
SD *Varireps l*	0.0187	0.0125	**0.0123**	**0.0115**	0.0087
eF *Variseq l*	0.0187	0.0125	**0.0123**	**0.0115**	0.0087
SD *Variseq l*	0.0187	0.0125	**0.0123**	**0.0115**	0.0087

Bold values denote improvements over the baseline

Table 6 NDCG@k for information need polyrepresentation

IN All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.0120	0.0131
eF $l = 5$	0.0068	0.0075	0.0082	0.0086	0.0090
SD $l = 5$	0.0068	0.0075	0.0082	0.0086	0.0090
eF $l = 10$	0.0068	0.0095	**0.0100**	**0.0125**	**0.0138**
SD $l = 10$	0.0068	**0.0097**	0.0097	**0.0127**	**0.0140**
eF *Varireps l*	0.0068	0.0095	0.0077	0.0081	0.0091
SD *Varireps l*	0.0068	**0.0097**	0.0075	0.0098	0.0109
eF *Variseq l*	0.0068	0.0095	0.0078	0.0078	0.0084
SD *Variseq l*	0.0068	**0.0097**	0.0099	0.0104	0.0111

Bold values denote improvements over the baseline

Table 7 P@k for document polyrepresentation

Doc All	P@5	P@10	P@15	P@20	P@30
BM25	0.1469	0.1375	0.1240	0.1117	0.1000
eF $l = 5$	**0.1500**	**0.1391**	**0.1292**	**0.1156**	0.0943
SD $l = 5$	**0.1500**	**0.1391**	**0.1292**	**0.1156**	0.0943
eF $l = 10$	**0.1500**	**0.1422**	**0.1302**	**0.1156**	0.0995
SD $l = 10$	**0.1500**	**0.1406**	**0.1292**	**0.1148**	0.0990
eF *Varireps l*	**0.1500**	**0.1422**	**0.1272**	0.1115	0.0964
SD *Varireps l*	**0.1500**	**0.1406**	0.1138	0.1038	0.0862
eF *Variseq l*	**0.1500**	**0.1422**	0.1128	0.1054	0.0836
SD *Variseq l*	**0.1500**	**0.1406**	0.1036	0.0862	0.0723

Bold values denote improvements over the baseline

Table 7 shows the P@k results for document polyrepresentation. Table 8 display the corresponding NDCG@k results. We can see some improvement over the baseline, in particular for our eF cluster ranking strategy in terms of NDCG, whereas the SD strategy often resulted in even worse results than the baseline.

The experiments confirm the trend that we already observed with the ideal clustering. We get higher values with slightly larger improvements for document polyrepresentation,

Table 8 NDCG@k for document polyrepresentation

Doc All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
eF $l = 5$	**0.0800**	**0.1076**	**0.1320**	**0.1461**	**0.1582**
SD $l = 5$	0.0607	**0.1076**	**0.1320**	**0.1461**	**0.1582**
eF $l = 10$	**0.0800**	**0.1089**	**0.1316**	**0.1445**	**0.1632**
SD $l = 10$	0.0607	0.0962	0.1189	0.1318	0.1318
eF *l=Varireps l*	**0.0800**	**0.1089**	**0.1314**	**0.1433**	**0.1601**
SD *l=Varireps l*	0.0607	0.0962	0.1036	0.1149	0.1264
eF *l=Variseq l*	**0.0800**	**0.1089**	**0.1233**	**0.1382**	0.1497
SD *l=Variseq l*	0.0607	0.0962	0.0962	0.1019	0.1112

Bold values denote improvements over the baseline

Table 9 High queries: P@k for IN polyrepresentation

IN High	P@5	P@10	P@15	P@20	P@30
BM25	0.0421	0.0316	0.0246	0.0263	0.0246
eF $l = 5$	**0.0632**	**0.0474**	**0.0386**	**0.0342**	**0.0263**
SD $l = 5$	**0.0632**	**0.0474**	**0.0386**	**0.0342**	**0.0263**
eF $l = 10$	**0.0632**	**0.0368**	**0.0351**	**0.0368**	**0.0316**
SD $l = 10$	**0.0632**	**0.0368**	**0.0351**	**0.0368**	**0.0316**
eF *Varireps l*	**0.0632**	**0.0368**	**0.0351**	**0.0316**	**0.0281**
SD *Varireps l*	**0.0632**	**0.0368**	**0.0316**	**0.0316**	**0.0298**
eF *Variseq l*	**0.0632**	**0.0368**	**0.0351**	0.0263	0.0175
SD *Variseq l*	**0.0632**	**0.0368**	**0.0386**	**0.0368**	**0.0281**

Bold values denote improvements over the baseline

whereas for information need polyrepresentation the results are mixed. Clearly, compared to the ideal ranking, there is room for improvement as none of the cluster-based results gained some significant effectiveness increase. However, we can also see that the approach nonetheless looks promising, in particular when it comes to document polyrepresentation. When it comes to IN polyrepresentation, it is interesting to observe that for instance for P@20, our cluster ranking approach sometimes delivers a marginally better result than the ideal cluster ranking. Given the overall low values for IN polyrepresentation, this might be just by chance, but it may be worth investigating. In any case it gives us a hint that more refined methods for cluster ranking to simulate the user behaviour are needed.

Results of proposed method (high and low)

One of the problems we faced with the iSearch collection is that some of the queries have a high number of relevant documents, while others only have very few documents judged relevant. We envisage that this has an effect on the performance of our proposed approach and investigate here its performance on 'hard' (<20 relevant documents) and 'easy' (20 and more relevant documents) queries. This way we identified 19 'easy' and 46 'hard' queries. We refer to the different sections as 'High' and 'Low'.

For IN polyrepresentation, P@k and NDCG@k scores for the High part of the evaluation are shown in Tables 9 and 10, respectively. For the queries with a high number of

Table 10 High queries: NDCG@k for IN polyrepresentation

IN High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0234	0.0243	0.0257	0.0270	0.0311
eF $l = 5$	0.0234	**0.0255**	**0.0281**	**0.0293**	0.0307
SD $l = 5$	0.0234	**0.0255**	**0.0281**	**0.0293**	0.0307
eF $l = 10$	0.0234	0.0243	**0.0261**	**0.0287**	**0.0332**
SD $l = 10$	0.0234	0.0147	0.0147	0.0185	0.0204
eF *Varireps l*	0.0234	0.0243	0.0173	0.0178	0.0186
SD *Varireps l*	0.0234	0.0147	0.0168	0.0171	0.0191
eF *Variseq l*	0.0234	0.0243	**0.0268**	0.0268	0.0268
SD *l*=*Variseq l*	0.0234	0.0147	**0.0277**	**0.0294**	**0.0316**

Bold values denote improvements over the baseline

relevant documents, we naturally get higher scores. It is also interesting to see that for these kinds of queries our approach provides some improvement at least in precision, which is an interesting result (although again no statistical significance can be reported here). There does, however, not seem to be much difference when it comes to the cluster ranking or document browsing strategy, except that $l = 5$ seems to be a good choice.

Tables 11 and 12 show the results for High queries and document polyrepresentation. Surprisingly, our clustering strategy does not seem to work well as no improvement over the baseline at all could be reported here.

Tables 13, 14, 15 and 16 show the results for considering the queries with a low number of relevant documents. We can clearly see improvements for document polyrepresentation and some lesser improvements for IN polyrepresentation (when it comes to NDCG). There are many 0 values due to the low number of relevant documents available for these queries. In particular for IN polyrepresentation, just selecting the top 5 documents per cluster ($l = 5$) suffers from the fact that there seem to be no relevant documents in the top 5, either in each cluster or in the overall baseline ranking. The situation is slightly better when it comes to document polyrepresentation, which seems to be capable of getting relevant documents into the top ranks both per cluster but also for the baseline ranking. It also seems our clustering strategy (in particular eF) is superior over a mere baseline ranking when it comes to queries with a low number of relevant documents. However, again we could not report statistical significance.

Table 11 High queries: P@k for document polyrepresentation

Doc High	P@5	P@10	P@15	P@20	P@30
BM25	0.3263	0.3000	0.2877	0.2447	0.2140
eF $l = 5$	0.3053	0.2895	0.2702	0.2395	0.2018
SD $l = 5$	0.3053	0.2895	0.2702	0.2395	0.2018
eF $l = 10$	0.3053	0.3000	0.2772	0.2447	0.2123
SD $l = 10$	0.3053	0.3000	0.2772	0.2447	0.2123
eF *Varireps l*	0.3053	0.3000	0.2737	0.2395	0.2140
SD *Varireps l*	0.3053	0.3000	0.2702	0.2447	0.2070
eF *Variseq l*	0.3053	0.3000	0.2316	0.2184	0.1789
SD *Variseq l*	0.3053	0.3000	0.2456	0.2053	0.1772

Table 12 High queries: NDCG@k for document polyrepresentation

Doc High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0708	0.1060	0.1308	0.1440	0.1688
eF $l = 5$	0.0708	0.1015	0.1264	0.1386	0.1568
SD $l = 5$	0.0708	0.1015	0.1264	0.1386	0.1568
eF $l = 10$	0.0708	0.1060	0.1308	0.1430	0.1638
SD $l = 10$	0.0708	0.1060	0.1308	0.1430	0.1430
eF *Varireps l*	0.0708	0.1060	0.1301	0.1421	0.1666
SD *Varireps l*	0.0708	0.1060	0.1008	0.1128	0.1292
eF *Variseq l*	0.0708	0.1060	0.1203	0.1342	0.1497
SD *Variseq l*	0.0708	0.1060	0.1160	0.1231	0.1395

Bold values denote improvements over the baseline

Table 13 Low queries: P@k for IN polyrepresentation

IN Low	P@5	P@10	P@15	P@20	P@30
BM25	0.0000	0.0022	0.0015	0.0022	0.0015
eF $l = 5$	0.0000	0.0000	0.0015	0.0000	0.0000
SD $l = 5$	0.0000	0.0000	0.0000	0.0000	0.0000
eF $l = 10$	0.0000	0.0022	0.0015	0.0022	0.0015
SD $l = 10$	0.0000	0.0022	0.0015	0.0022	0.0015
eF *Varireps l*	0.0000	0.0022	0.0001	0.0001	0.0001
SD *Varireps l*	0.0000	0.0022	0.0001	0.0011	0.0007
eF *Variseq l*	0.0000	0.0022	0.0001	0.0001	0.0007
SD *Variseq l*	0.0000	0.0022	0.0015	0.0011	0.0007

Table 14 Low queries: NDCG@k for IN polyrepresentation

IN Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0000	0.0034	0.0034	0.0057	0.0057
eF $l = 5$	0.0000	0.0000	0.0000	0.0000	0.0000
SD $l = 5$	0.0000	0.0000	0.0000	0.0000	0.0000
eF $l = 10$	0.0000	0.0034	0.0034	**0.0058**	**0.0058**
SD $l = 10$	0.0000	**0.0076**	**0.0076**	**0.0103**	**0.0114**
eF *Varireps l*	0.0000	0.0034	0.0034	0.0041	0.0052
SD *Varireps l*	0.0000	**0.0076**	**0.0037**	**0.0067**	**0.0076**
eF *Variseq l*	0.0000	0.0034	0.0001	0.0001	0.0007
SD *Variseq l*	0.0000	**0.0076**	0.0026	0.0026	0.0026

Discussion

Based on the observation that both polyrepresentation and clustering create a partitioning of the document set, we have evaluated several cluster-based exploration strategies for polyrepresentation to a polyrepresentative baseline. By applying a kind of ideal cluster ranking we have demonstrated that the general strategy indeed bears the potential for a

Table 15 Low queries: P@k for document polyrepresentation

Doc Low	P@5	P@10	P@15	P@20	P@30
BM25	0.0800	0.0689	0.0593	0.0544	0.0489
eF $l = 5$	**0.0844**	**0.0756**	**0.0696**	**0.0633**	0.0489
SD $l = 5$	**0.0844**	**0.0756**	**0.0696**	**0.0633**	0.0489
eF $l = 10$	**0.0844**	**0.0756**	**0.0681**	**0.0611**	**0.0600**
SD $l = 10$	0.0800	**0.0733**	**0.0667**	**0.0600**	**0.0511**
eF *Varireps l*	**0.0844**	**0.0756**	**0.0667**	**0.0587**	0.0478
SD *Varireps l*	0.0800	0.0733	0.0493	0.0457	0.0362
eF *Variseq l*	**0.0844**	**0.0756**	**0.0638**	**0.0587**	0.0442
SD *Variseq l*	0.0800	0.0522	0.0449	0.0370	0.0290

Bold values denote improvements over the baseline

Table 16 Low queries: NDCG@k for document polyrepresentation

Doc Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0771	0.0993	0.1167	0.1316	0.1520
eF $l = 5$	**0.0839**	**0.1101**	**0.1343**	**0.1492**	**0.1588**
SD $l = 5$	**0.0839**	**0.1101**	**0.1343**	**0.1492**	**0.1588**
eF $l = 10$	**0.0839**	**0.1101**	**0.1319**	**0.1451**	**0.1630**
SD $l = 10$	**0.0839**	0.0921	0.1139	0.1272	0.1272
eF *Varireps l*	**0.0839**	**0.1101**	**0.1320**	**0.1439**	**0.1575**
SD *Varireps l*	**0.0839**	0.0921	0.0988	0.1102	0.1205
eF *Variseq l*	**0.0839**	**0.1101**	**0.1245**	**0.1398**	0.1496
SD *Variseq l*	**0.0839**	0.0921	0.0881	0.0932	0.0996

Bold values denote improvements over the baseline

more effective search experience. Applying several cluster ranking strategies along with document exploration ones showed improvements on the one hand, but also that our model needs to be refined to eventually get statistically significant results. A reason for not getting significant improvements could be the overly simple user model (SD/eF for cluster ranking, $l = 5, 10$ and *Varireps l* and *Variseq l* for exploring documents within clusters) that we applied in our evaluation to get an artificial ranking that can be compared to a baseline ranking. Our assumptions for simulated users are very basic and focus on a very simple objective of interaction. While these strategies can be applied in systems that present their users with a ranked list of documents, the main motivation of the simulated user approach is indeed that users themselves decide which cluster to choose next (and we just try to model how this decision could be made for evaluation purposes). In a real scenario, users may know better than our proposed algorithms which cluster to choose next, and that may lead to improvements that may even exceed what we deemed an ideal cluster ranking in this paper. The hypothesis thus is that our approach, applied in a system that lets users explore clusters based on polyrepresentation, will eventually support the user better than any system offering just one linear ranked result list. Shedding some light into this of course implies that we leave our simulated user framework in favour of a 'real' user study.

Our study also reveals some further interesting insight regarding the difference between information need and document polyrepresentation. While overall document polyrepresentation, which is also exploiting bibliographic evidence like citations, seems to be the preferred choice over information need polyrepresentation, we found a different picture when it comes to 'hard' and 'easy' queries. Document polyrepresentation that considers citations seem to work better on queries with a low number of relevant documents ('hard' queries) while information need polyrepresentation clustering, though still producing low scores, was able to beat the baseline for queries with a high number of relevant documents ('easy' queries). However, we need to bear in mind that the number of 'easy' queries is quite low (19), which may have influenced the results. Further investigation, also with different kinds of representations, is required to confirm this finding. While the information need based polyrepresentation at first glance does not seem to have a connection to scientometrics, our motivation of reporting such experiments is directly related to the scientific literature search where the information need is also a crucial component to understand and to model. This is why the polyrepresentation of information needs is widely used in the literature exploring cognitive models.

From the discussion so far it could be inferred from the simulated user approach that the top ranked clusters in the ideal scenario have many relevant documents. Thus, if the clusters are ranked nearer to the ranking created in the ideal scenario then the performance could significantly be improved as compared to the baseline. In many but not all cased the eF measure shows some improvement both in IN and document based polyrepresentation SD.

Conclusion and future work

In this paper we presented ways to utilise polyrepresentative partitions with a document clustering approach. The assumption is that in a system supporting polyrepresentation, clustering can overcome the problem that the system does not know about the user's preference regarding the representations he or she deems important. Based on the OCF we introduced clustering to polyrepresentation of information needs and documents, the latter utilising bibliometric means like citations. In our evaluation we simulated the user by applying different cluster ranking and within-cluster document selection strategies to create a ranking that reflects the documents a user would investigate throughout the process. The evaluation pointed out some interesting insights and open areas for further research. In particular we have demonstrated in a somewhat ideal scenario that the basic idea may indeed lead to statistically significant results. We have then shown how different cluster ranking strategies can contribute to a solution. We further found that information need polyrepresentation seems to be a better choice for queries with less relevant documents, while document polyrepresentation seems to work well with queries having many relevant documents.

While this study produced some very interesting insights also into the nature of polyrepresentation in the context of the iSearch collection, many questions are still open. The context in this study for a document was derived from the title and abstracts of the documents cited in there. The other option could be to drive the context for a document from the documents it is cited in. Besides this picking the actual 50–100 words around the actual citation point could be another option to create the citation context. The cluster ranking is a second point which needs more sophisticated definition of the ranking function, although the eF and SD measures used here show some performance improvement.

The assumptions made for the simulated user are very basic. To further display a proof of concept, the simulated user model could be improved in several ways. For instance, we may look into using several values (in a certain range) for choosing the l documents from the clusters and the sample of such users could be employed to make the simulation more realistic. Besides this, the actual user search behaviour could be used to improve the simulated user. In general we plan to look into the possibility of performing a more complete simulation study of the different paths a user might take when exploring clusters and within-cluster rankings, which could lead to a more refined theoretical upper bound achievable by our approach. Another improvement could be labelling the clusters with few most prominent key phrases extracted from within the cluster so that the simulated user matches the information need (query) with each cluster label to guide the search process. Besides this the representation information could also be used within the label to indicate the belonging of the cluster to the particular representation, which could as well be used to guide the search process. As we have observed a different behaviour of information need and document polyrepresentation, respectively, when it comes to 'hard' and 'easy' queries, our plan is to explore a combination of IN and document polyrepresentation as indicated in Eq. 4. Another part to investigate addresses our choice of using BM25 for approximating the probability of relevance. Our plan is to look at the effect of different scoring functions, in particular language models, on the effectiveness of the approach. Furthermore we will look into methods that provide a more direct estimation of the probability of relevance, for instance as described by (Nottelmann and Fuhr 2003).

Acknowledgments The authors would like to thank anonymous reviewers for their insightful comments.

References

Azzopardi, L. (2011). The economics in interactive information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval (SIGIR 2011)* (pp. 15–24). New York, New York, USA: ACM Press.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multipleperspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386–1409.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In N. Belkin, P. Ingwersen, A. M. Pejtersen & E. A. Fox (Eds.), *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR'92* (pp. 318–329).

Fox, E. A., & Shaw, J. A. (1993). Combination of multiple searches. In D. Harman (Ed.), *The second text retrieval conference (TREC-2)*. National Institute of Standards and Technology, Gaithersburg, Md. 20899 (pp. 243–252).

Frommholz, I., & Abbasi, M. K. (2014). On clustering and polyrepresentation. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky & K. Hofmann (Eds.), *Proceedings of the European conference on information retrieval (ECIR 2014)* (Vol. 1, pp. 618–623). Springer.

Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., & van Rijsbergen, K. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceedings of the 2010 information interaction in context symposium* (pp. 115–124). New Brunswick: ACM.

Fuhr, N., Lechtenfeld, M., Stein, B., & Gollub, T. (2011). The optimum clustering framework: Implementing the cluster hypothesis. *Information Retrieval*, *15*(2), 93–115.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005a). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548–1572.

Glenisson, P., Glänzel, W., & Persson, O. (2005b). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, *63*(1), 163–180.

Goldszmidt, M., & Sahami, M. (1998). *A probabilistic approach to full-text document clustering*. Technical report ITAD-433-MS-98-044, SRI International.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the SIGIR 1996* (pp. 76–84).

Hull, D. (1993) Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 329–338). ACM.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *The Journal of Documentation*, *52*(1), 3–50.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Secaucus, NJ, USA: Springer

Ji, X., & Xu, W. (2006). Document clustering with prior knowledge. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 405–412). ACM.

Ke, W., & Sugimoto, C. R., & Mostafa J. (2009). Dynamicity vs. effectiveness: Studying online clustering for scatter/gather. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 19–26). ACM.

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, *43*(1), 30–46.

Larsen, B. (2002). Exploiting citation overlaps for information retrieval: Generating a boomerang effect from the network of scientific papers. *Scientometrics*, *54*(2), 155–178.

Larsen, B., Ingwersen, P., & Kekäläinen, J. (2006). The polyrepresentation continuum in IR. In *IIiX: Proceedings of the 1st international conference on information interaction in context* (pp. 88–96). New York, NY, USA: ACM.

Larsen, B., Lioma, C., Frommholz, I., & Schütze, H. (2012). Preliminary study of technical terminology for the retrieval of scientific book metadata records categories and subject descriptors. In *SIGIR 2012: Proceedings of the 35th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 1131–1132).

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. In *Proceedings of the tenth international conference on information and knowledge management* (pp. 33–40). ACM.

Lüke, T., Schaer, P., & Mayr, P. (2013). A framework for specific term recommendation systems. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval—SIGIR'13* (p. 1093).

Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, R. Stefan & K. Rijsbergen (Eds.), *Proceedings ECIR 2010* (pp. 627–630). Berlin, Heidelberg: Springer.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, California, USA* (Vol. 1).

Mayr, P., & Mutschke, P. (2013). Bibliometric-enhanced retrieval models for big scholarly information systems. In *Proceedings IEEE international conference on big data workshop on scholarly big data: Challenges and ideas*.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. *Library Review*, *57*(3), 213–224.

Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics*, *89*(1), 349–364.

Na, S.-H., Kang, I.-S., & Lee, J.-H. (2007). Adaptive document clustering based on query-based similarity. *Information Processing & Management*, *43*(4), 887–901.

Nottelmann, H., & Fuhr, N. (2003). From retrieval status values to probabilities of relevance for advanced IR applications. *Information Retrieval*, *6*(4), 363–388.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma C. (2006). Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 workshop on open source information retrieval (OSIR 2006)*.

Raiber, F., & Kurland, O. (2013). Ranking document clusters using markov random fields. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval-SIGIR'13* (pp. 333–342).

Ritchie, A., Teufel, S., & Robertson, S. (2006) Creating a test collection for citation-based IR experiments. In *Proceedings of the human language technology conference of the NAACL, main conference*.

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1998). Okapi at TREC-7. In *Proceedings of the 7th text retrival converence (TREC-7)*.

Schaer, P., Mayr, P., & Lüke, T. (2012). Extending term suggestion with author names. In *Proceedings of theory and practice of digital libraries 2012 (TPDL 2012)*.

Skov, M., Larsen, B., & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing & Management*, *44*(5), 1673–1683.

Smucker M. D., Allan J., Carterette B. (2007) A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM conference on information and knowledge management (CIKM)* (pp. 623–632). ACM.

Tombros, A., Villa, R., & Van Rijsbergen, C. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management*, *38*(4), 559–582.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.

Webber, W., Moffat, A., Zobel, J., & Sakai T. (2008). Precision-at-ten considered redundant. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 695–696). ACM.