Polyrepresentative Clustering: A Study of Simulated User Strategies and Representations

Muhammad Kamran Abbasi    Ingo Frommholz

Institute for Research in Applicable Computing
University of Bedfordshire UK

Second BIR Workshop, ECIR2015
29 March 2015
Outline

1. Introduction
2. Polyrepresentation and Clustering
3. Evaluation
4. Conclusion
Introduction

- Principle of Polyrepresentation in IIR
- Multiple representations of information need and information object (documents)
  - Cognitive overlap supposed to contain relevant documents
- Combination of document clustering and polyrepresentation
Polyrepresentation and Clustering

- Polyrepresentation creates partitions
- Clustering partitions document sets too
- Can clustering help in creating polyrepresentative partitions?
Information Need-based Vector

- Let $REP_{in}$ be the set of representations\(^1\) of an information need $in$
- Motivated by the Optimum Clustering Framework (OCF) which is based on the probability of relevance (Fuhr et al., 2011)
- $\Pr(R|d, r_i)$ is computed for each document $d$ and $r_i \in REP_{in}$

\[
\vec{\tau}_{in}(d) = \begin{pmatrix}
\Pr(R|d, r_1) \\
\vdots \\
\Pr(R|d, r_n)
\end{pmatrix}
\]

\(^1\)search terms, work task, ideal answer, current info need, background knowledge
Document-based Polyrepresentation Vector

- $REP_d$ consists of the different representations $^2rd_i$ of a document $d$
- thus the $Pr(R|rd_i, q)$ for $q$ (search terms in this case) is computed

$$
\vec{\tau}_{doc}(d) = \begin{pmatrix}
Pr(R|rd_1, q) \\
\vdots \\
Pr(R|rd_n, q)
\end{pmatrix}
$$

$^2$title, abstract, body, context, references
Bibliographic context
Representation Concatenation and Combinations

IN and Doc representation concatenation and combinations were used.

For example:

- **Concatenation of** $REP_{doc} \mid REP_{in}$:

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

- **Combination of** $REP_{doc}$ or $REP_{in}$
  - for Doc: {title, abstract}, {title, body text}...
  - for IN: {search terms, work task}, {search terms, ideal answer}...
Simulated User Strategies

- **Simulated User Strategy-1**
  - From each cluster: take top $l$ documents (sorted based on weights) and add them to a list
  - Sort documents in final list based on their weights and evaluate

- **Simulated User Strategy-2**
  - From first cluster take 1st document, add it to the list
  - Check if this document is relevant, if it is, then take next document from same cluster
  - If added document is not relevant switch to next cluster and take its first document
  - Follow the procedure until last cluster is reached
  - Sort documents in final list based on their weights and evaluate
Experiment Setup

- PF (full text) sub collection of iSearch collection
  - 65 search tasks
- IN and Document vectors as discussed above
- Terrier 3.5 was used for indexing and retrieval
- Using k-means $2^{\left|REP\right|}$ number of cluster were computed
- BM25 to estimate $Pr(R|rd_i, q)$ and $Pr(R|d, r_n)$, then apply Strategy-1 resp. Strategy-2 (yields a ranking)
- Baseline BM25 ranking: CombSUM of $Pr(R|rd_i, q)$ and $Pr(R|d, r_n)$ (its respective concatenation and combination)
Evaluation Results
Strategy 1 & 2 for IN & Doc Reps Concatenated

* shows statistically significant difference from baseline at $p < 0.05$
Evaluation Results

Strategy 1 for Doc Rep combination (title abstract)

![Bar chart showing NDCG@5 and NDCG@10 for Strategy-1 and BM25]
Evaluation Results

Strategy 1 for Doc Rep combination (abstract context)

**Strategy-1: Document Rep combinations (abstract context)**

![Graph showing NDCG@5 and NDCG@10 for Strategy-1 and BM25]

- **NDCG@5**: Strategy-1 = 0.08, BM25 = 0.06
- **NDCG@10**: Strategy-1 = 0.09, BM25 = 0.07
Evaluation Results

Strategy 1 for Doc Rep combination (context reference)

![Graph showing NDCG@5 and NDCG@10 for Strategy 1 with BM25 and arithMean combinations]

Strategy-1: Document Rep combinations (context Ref)

- **NDCG@5**
  - BM25: 0.05
  - arithMean: 0.06

- **NDCG@10**
  - BM25: 0.07
  - arithMean: 0.08
A polyrepresentative clustering strategy seems to improve effectiveness

Bibliometric information i.e. citation context and references could be helpful as representations (but needs further investigation)

(Simulated) user strategies have potential to be used for Interactive IR evaluation