Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering

Kamran Abbasi[1] Ingo Frommholz[1]

[1]Institute of Research in Applicable Computing
University of Bedfordshire UK

BIR Workshop, ECIR2014
13 April 2014

Outline

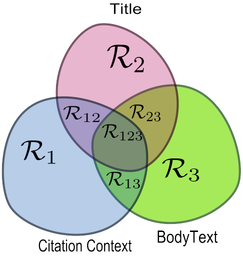
1. Introduction
2. Polyrepresentation and Clustering
3. Evaluation
4. Conclusion

Introduction

- Principle of Polyrepresentation in IIR
- Representations of information need and information object
- Helps to minimize the gap between user's space and information space
- Document Clustering

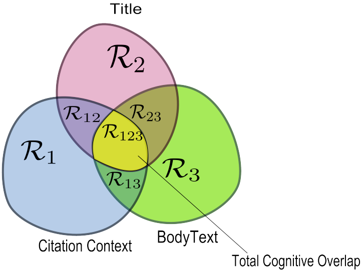
Polyrepresentation and Clustering

- Mapping of clusters to polyrepresentation
- Search strategy:
  1. User investigates total cognitive overlap cluster
  2. User jumps to different cluster based on preferences



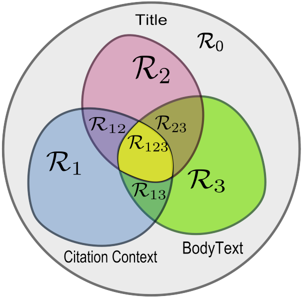
Polyrepresentation and Clustering

- Mapping of clusters to polyrepresentation
- Search strategy:
  1. User investigates total cognitive overlap cluster
  2. User jumps to different cluster based on preferences



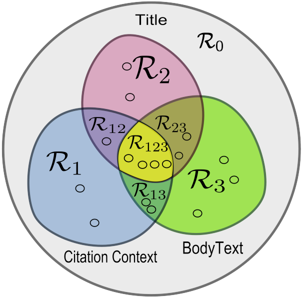
Polyrepresentation and Clustering

- Mapping of clusters to polyrepresentation
- Search strategy:
  1. User investigates total cognitive overlap cluster
  2. User jumps to different cluster based on preferences



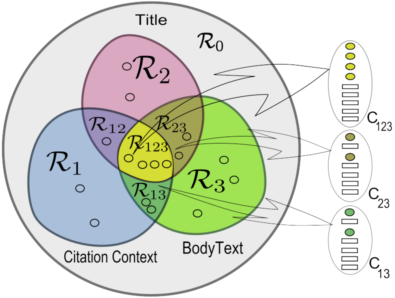
Polyrepresentation and Clustering

- Mapping of clusters to polyrepresentation
- Search strategy:
  1. User investigates total cognitive overlap cluster
  2. User jumps to different cluster based on preferences



Polyrepresentation and Clustering

- Mapping of clusters to polyrepresentation
- Search strategy:
  1. User investigates total cognitive overlap cluster
  2. User jumps to different cluster based on preferences



Simulated User and Cluster-based Ranking

- Rough simulation of search strategy
- Creates a ranking that we evaluate against baseline

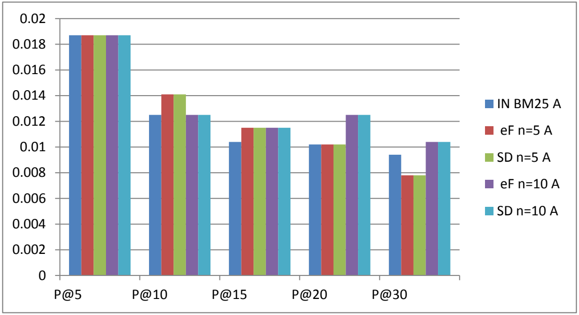
```
Require: Clustering $\mathcal{C}$, $k$
  $r \leftarrow ()$ {The ranking, initially an empty list}
  $C \leftarrow$ ranked list of clusters in $\mathcal{C}$ (using eF or SD)
  for all cluster $c \in C$ do
    $l \leftarrow$ ranked list of documents in $c$ {process $C$ in descending weight order}
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

Experiment Setup

- PF (full text) sub collection of iSearch collection
- Collection's citation information is used for context extraction
- Terrier3.5 Search Engine

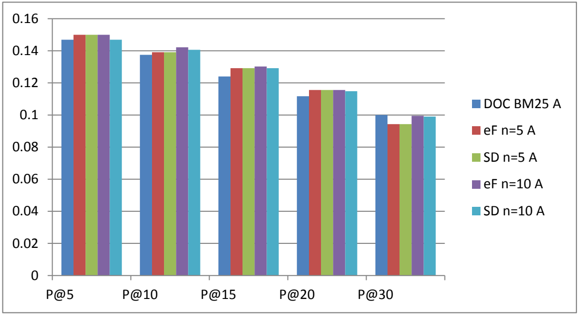
Evaluation Results

Results for All Queries IN



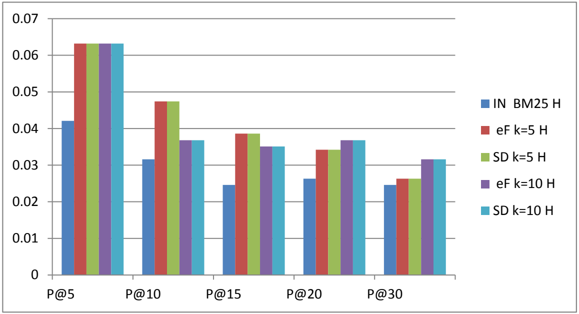
Evaluation Results

Results for All Queries Doc



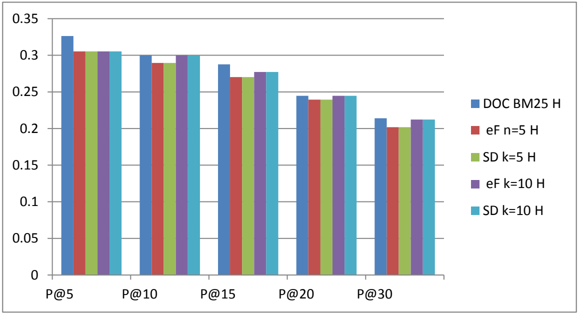
Evaluation Results

Queries with High Relevance Information IN



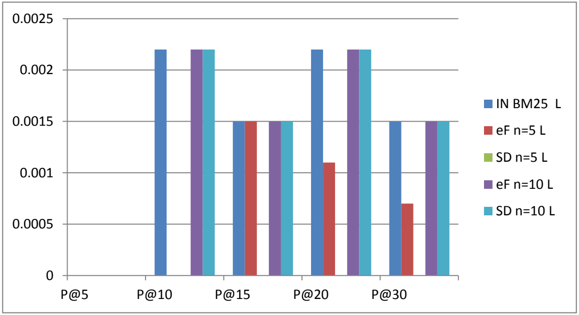
Evaluation Results

Queries with High Relevance Information Doc



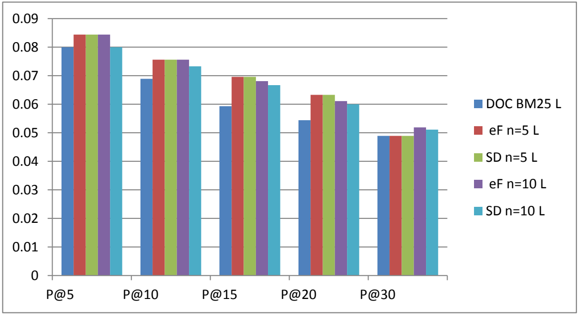
Evaluation Results

Queries with Low Relevance Information IN



Evaluation Results

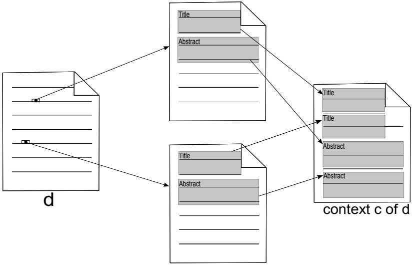
Queries with Low Relevance Information Doc



Conclusion

- Cluster ranking and cluster-based ranking have potential
- Bibliometric information i.e citation context and references show improvement on IR performance when combined with clustering
- Simulated user based evaluation of interactive systems can be enhanced

Bibliographic context


Title
Abstract
Title
Abstract
d
Title
Title
Abstract
Abstract
context c of d