# Contextual Suggestion Track

# TREC

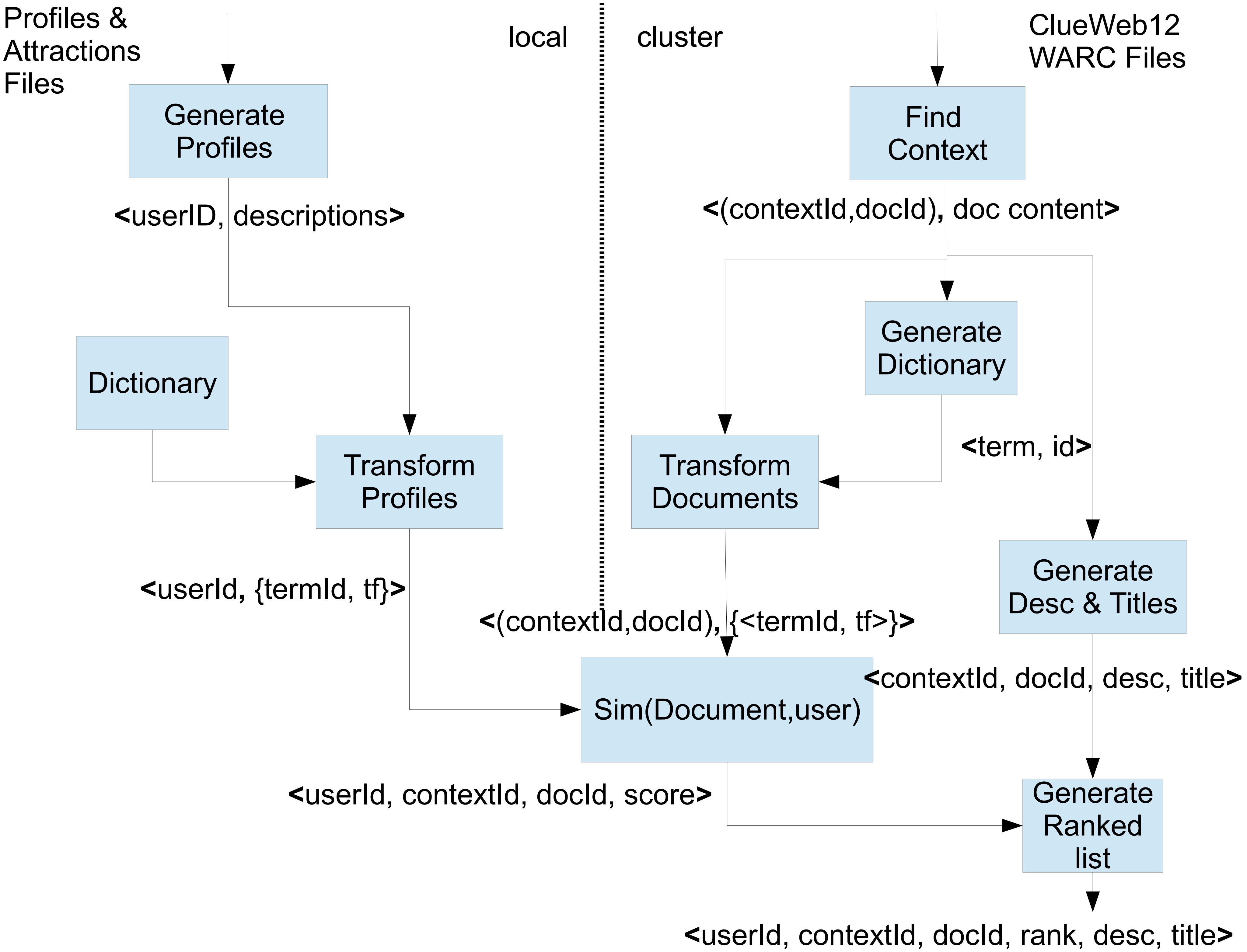
Thaer Samar, Alejandro Bellogin,
Jimmy Lin, Arjen P. de Vries, Alan Said

# Summary

- Content-based recommendation
  - Computes the similarity between documents and users profiles
- Classifier (not submitted)
  - Training data:
    - + Yelp, tripadvisor, wikitravel, zagat, yahoo-travel, orbitz
    - - Random sample
- Using full ClueWeb12

# ClueWeb12

- Statistics:
  - From February to May 2012
  - 5.5 TB (compressed)
  - 27.3 TB (uncompressed)
  - 33,447 WARC files
  - 733,019,372 documents
- Hadoop cluster:
  - 90 computing nodes
  - 720 parallel map/reduce tasks


Profiles & Attractions Files
local
cluster
ClueWeb12 WARC Files
Generate Profiles
Find Context
<userID, descriptions>
<(contextId,docId), doc content>
Generate Dictionary
Dictionary
Transform Profiles
Transform Documents
<term, id>
Generate Desc & Titles
<userId, {termId, tf}>
<(contextId,docId), {<termId, tf>}>
Sim(Document,user)
<contextId, docId, desc, title>
<userId, contextId, docId, score>
Generate Ranked list
<userId, contextId, docId, rank, desc, title>

# Find Context

- Goal: extract relevant documents for each context

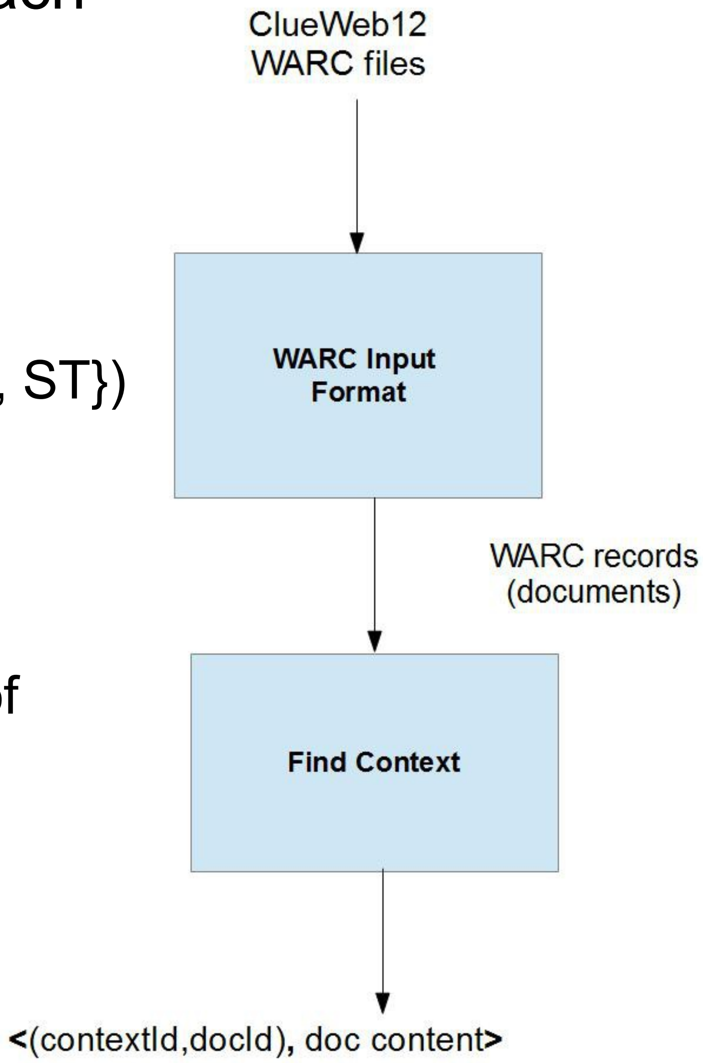
- How do we measure the relevance?
  - Exact mention of the context (format: {City, ST})

    Kennewick, WA
  - Exclude non related sentences

    I am in Kennewick, washing ...
  - Exclude documents that mention the city of interest but in different states

    Greenville, NC and Greenville, SC

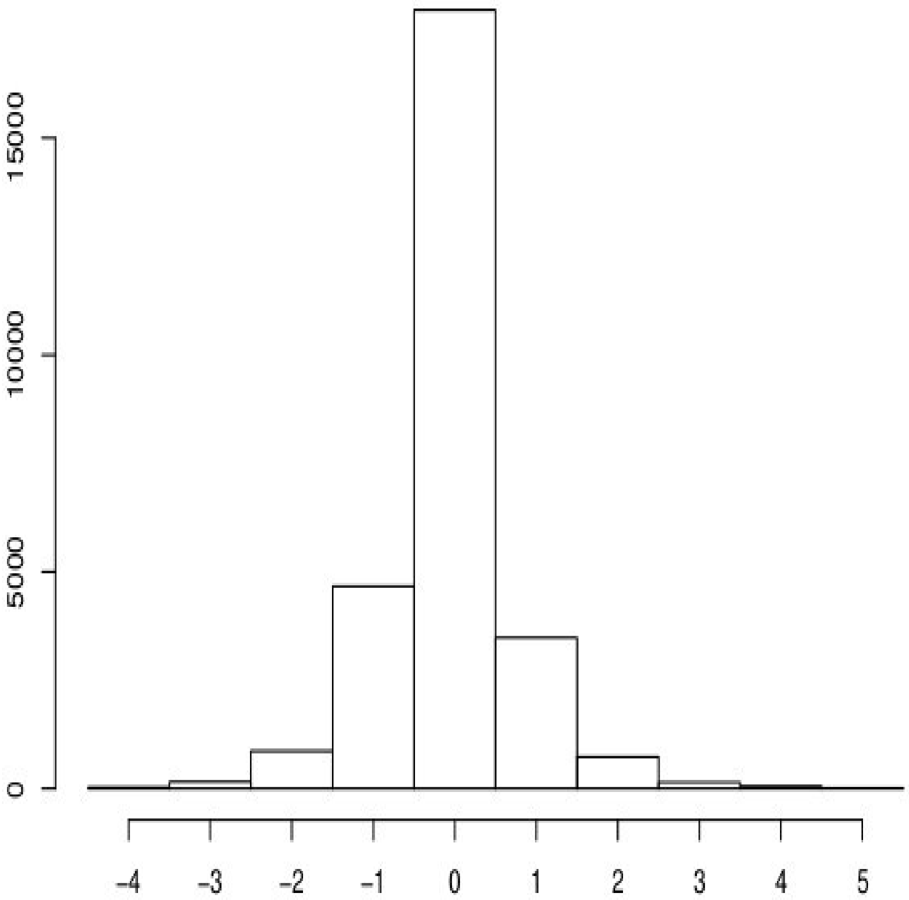
- We found 13,548,982 documents out of 733,019,372 ClueWeb12 documents



# Generate profiles

- We used the description of attractions rated by the user to generate his profile

- Why descriptions not the attraction website
  - 7 urls were found with one-one matching
  - 35 were found considering hostname matches and url variation, .i.e, http(s), www
  - ratings for the attraction's descriptions and websites were very similar

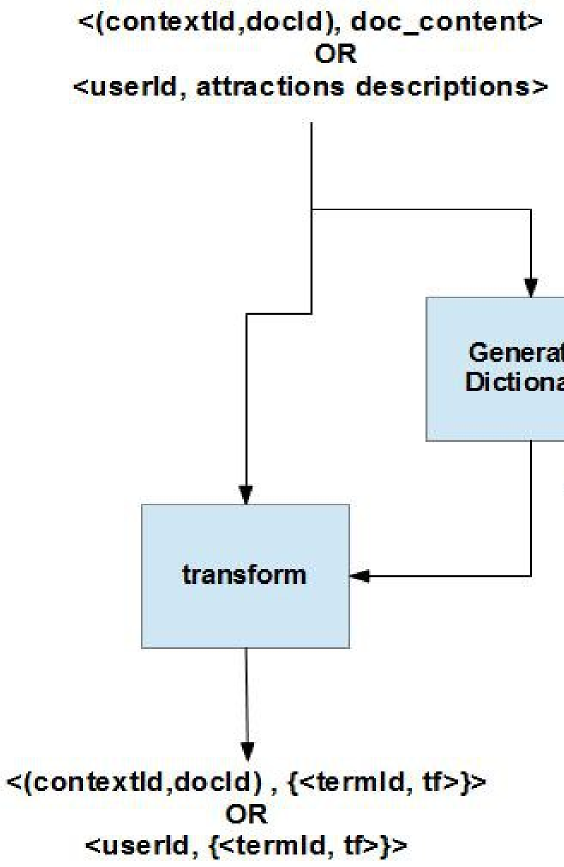


# Documents & profiles representation

- Vector Space Model
- Elements of the vectors are <term, frequency> pairs
- Efficient in terms of:
  - Size

    918 GB (before)

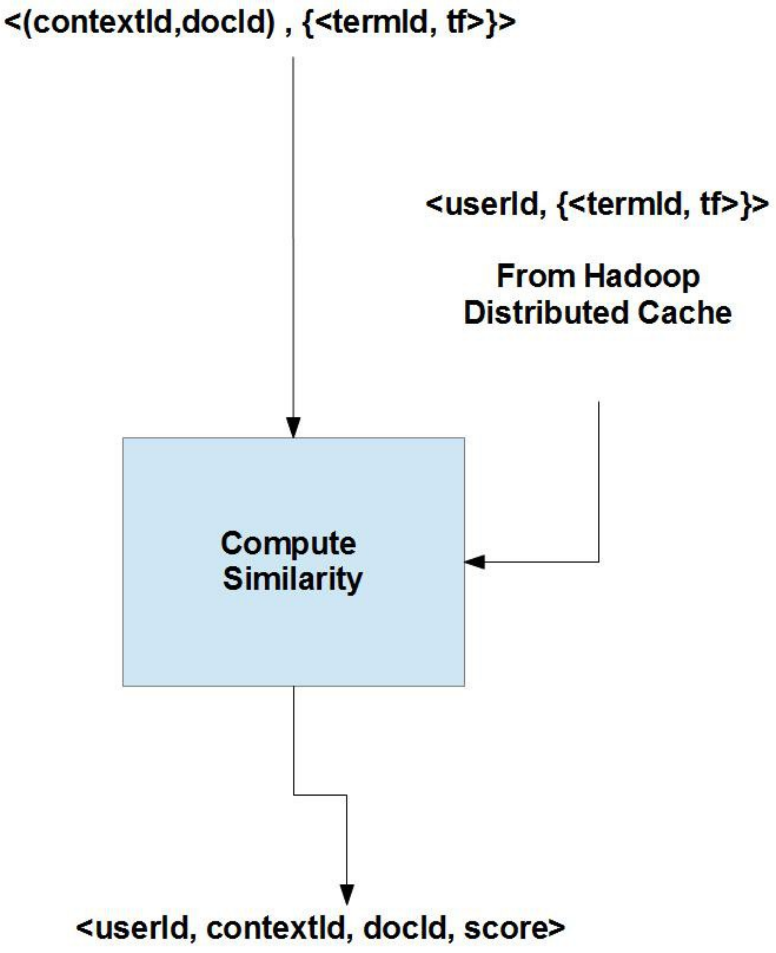
    40 GB (after)
  - Processing speed
- More complete implementation in https://github.com/lintool/clueweb



<(contextId,docId), doc_content>
OR
<userId, attractions descriptions>
Generate Dictionary
<term, id>
transform
<(contextId,docId) , {<termId, tf>}>
OR
<userId, {<termId, tf>}>

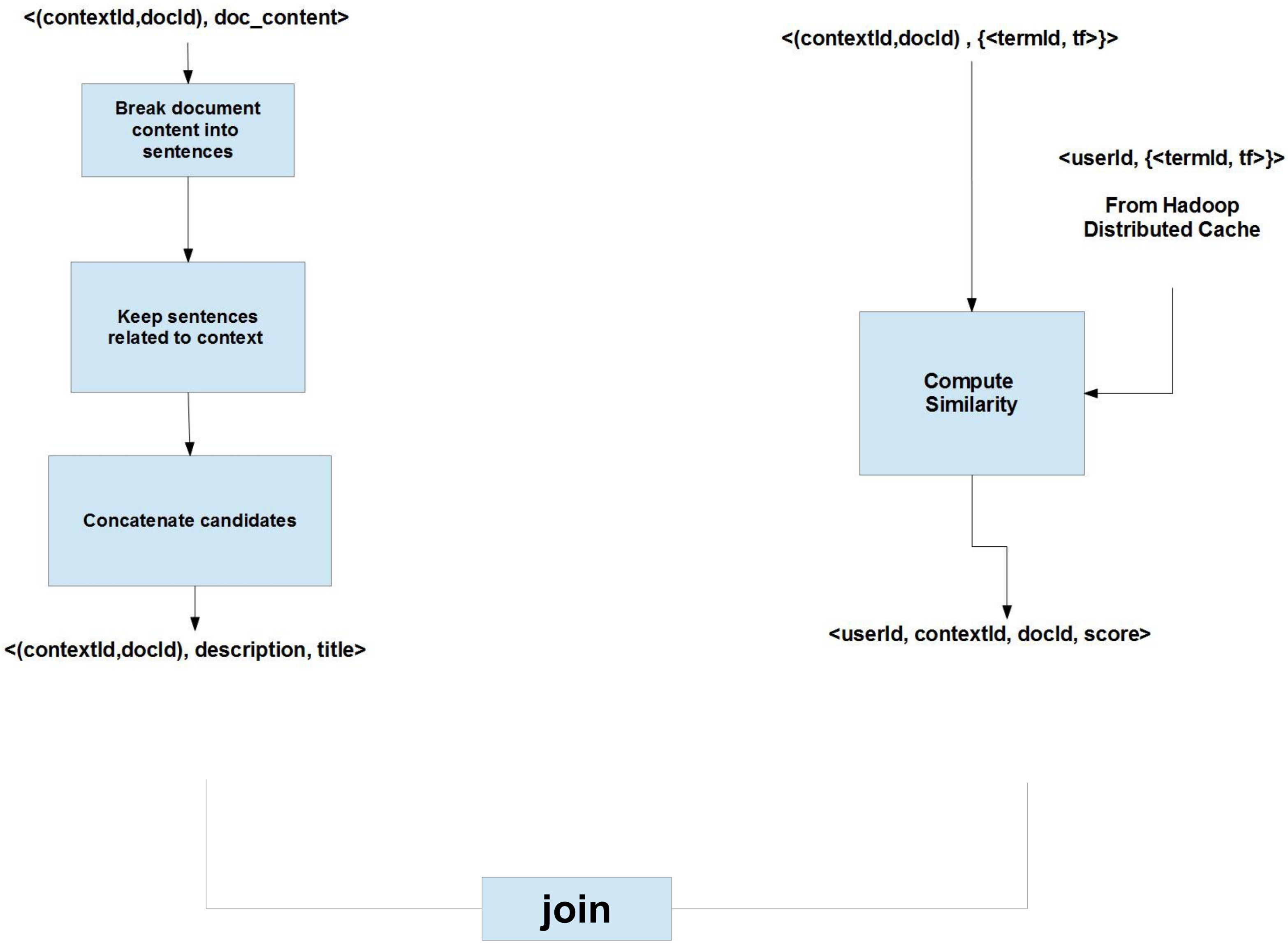
# Similarity

- Cosine similarity between profile and document vector space representation



<(contextId,docId) , {<termId, tf>}>
<userId, {<termId, tf>}>
From Hadoop Distributed Cache
Compute Similarity
<userId, contextId, docId, score>

# Descriptions and final results



<(contextId,docId), doc_content>

Break document content into sentences

Keep sentences related to context

Concatenate candidates

<(contextId,docId), description, title>

<(contextId,docId) , {<termId, tf>}>

<userId, {<termId, tf>}>

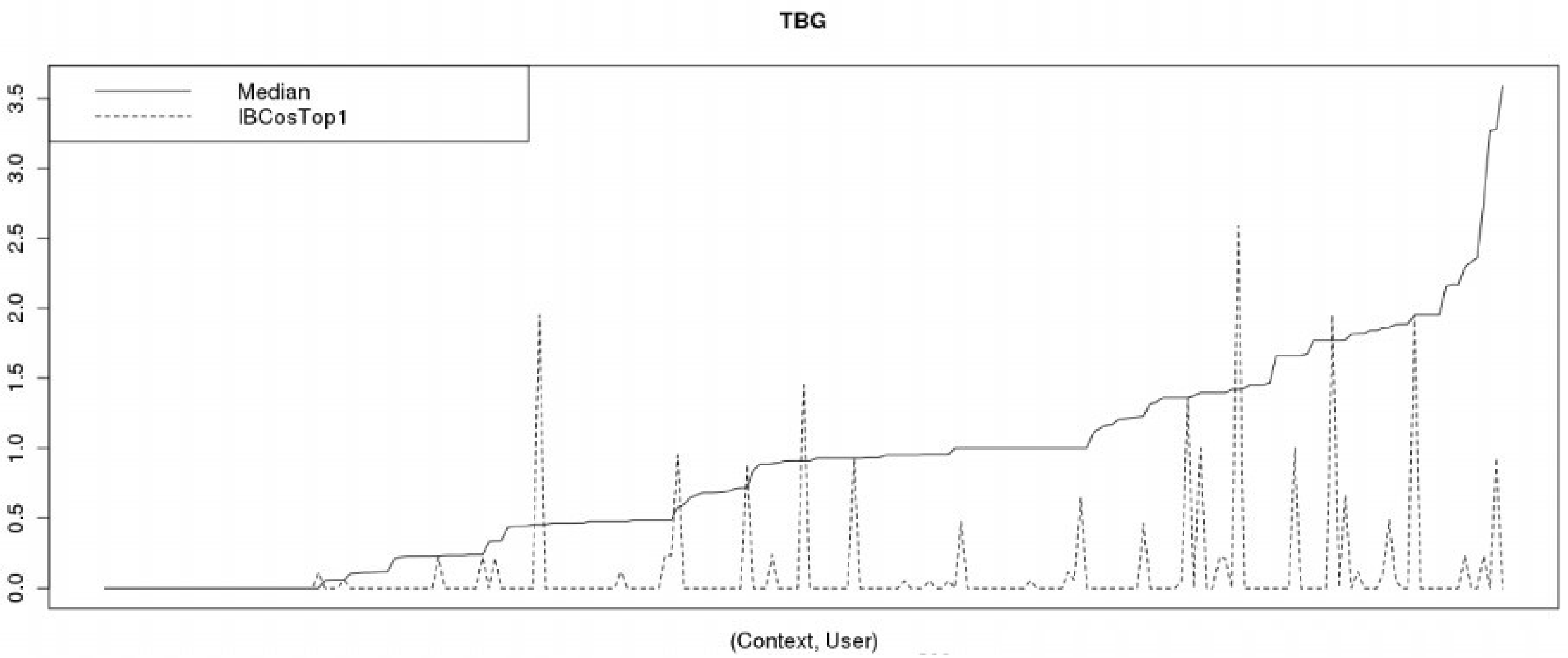
From Hadoop Distributed Cache

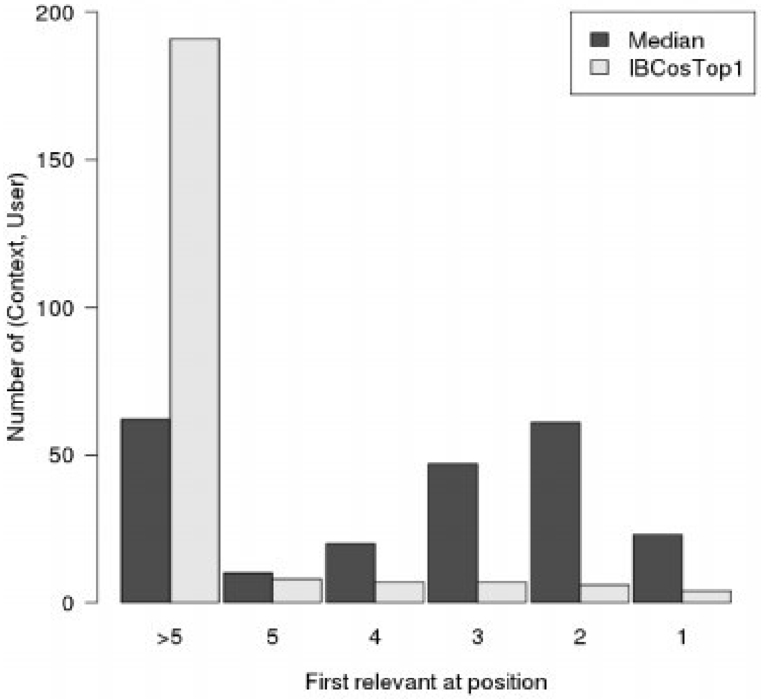
Compute Similarity

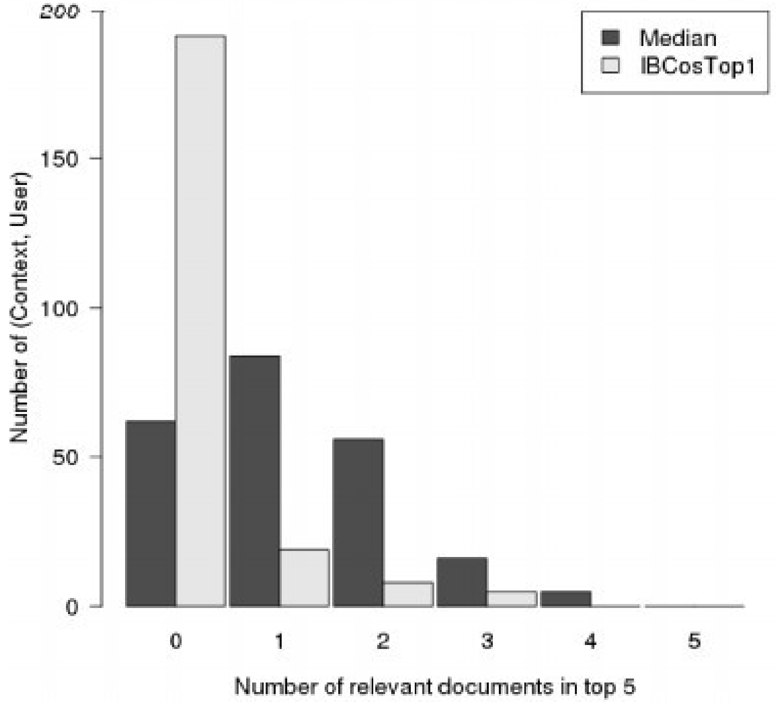
<userId, contextId, docId, score>

join

# Results


TBG
Median
IBCosTop1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
(Context, User)


Median
IBCosTop1
200
150
100
50
0
Number of (Context, User)
>5
5
4
3
2
1
First relevant at position

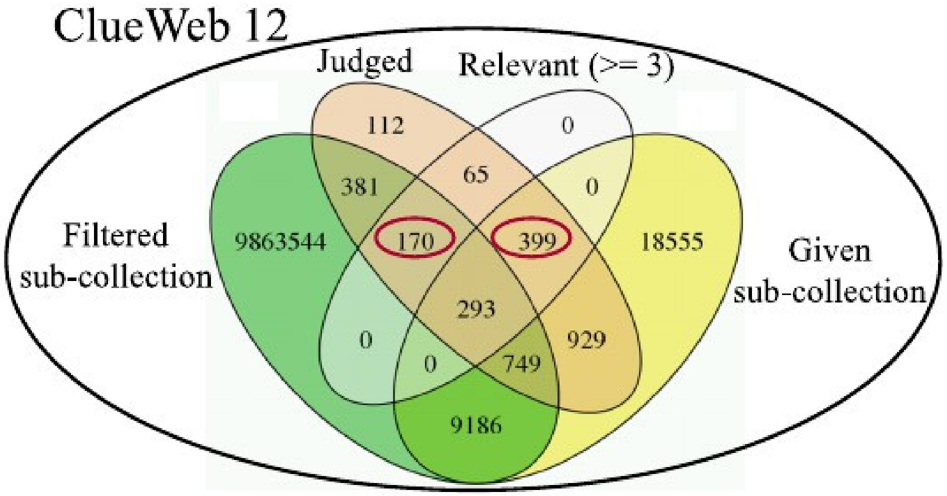

Median
IBCosTop1
200
150
100
50
0
Number of (Context, User)
0
1
2
3
4
5
Number of relevant documents in top 5

# Analysis

- We asked the following questions
  - Effect of sub-collection creation (context finding)
  - Effect of similarity function
  - Rating bias in ClueWeb vs Open Web

# Effect of sub-collection creation 1/2

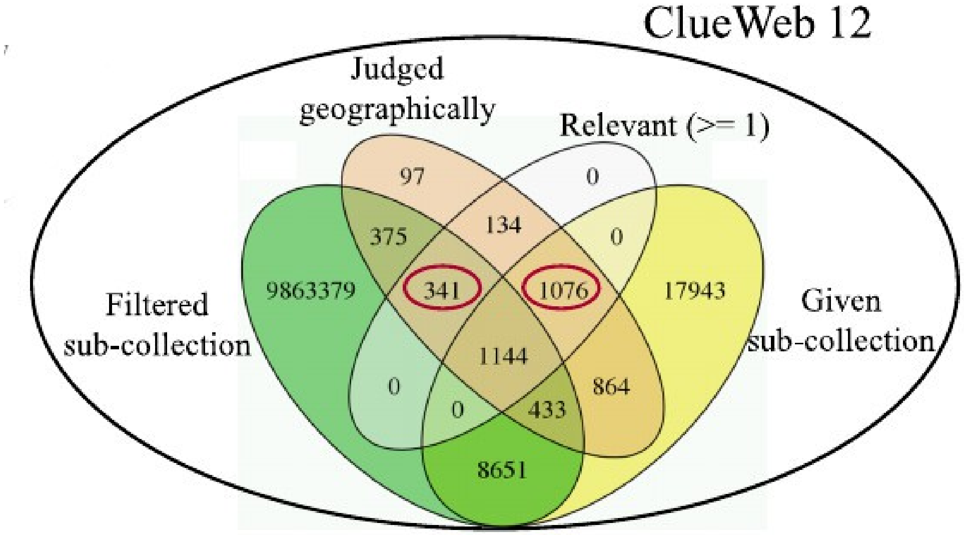
- Re-run our approach on the sub-collection given by organizers
  - 27% of given sub-collection are in our sub-collection

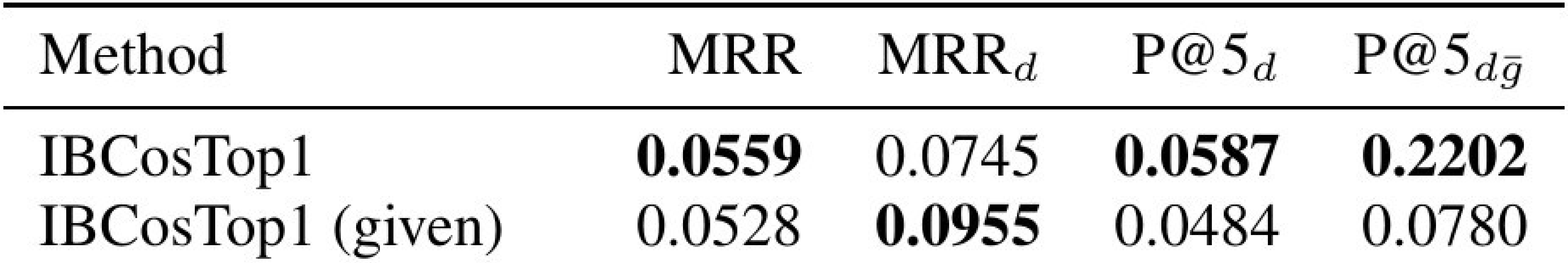


| Method | MRR | MRR$_d$ | P@5$_d$ |
|---|---|---|---|
| IBCosTop1 | **0.0559** | 0.0745 | **0.0587** |
| IBCosTop1 (given) | 0.0528 | **0.0955** | 0.0484 |

# Effect of sub-collection creation 2/2

- Significant improvement when ignoring the geographical aspect (P@5_g)
- Our method retrieves relevant documents for the user but not geographically appropriate
- The given sub-collection is more appropriate for the contexts

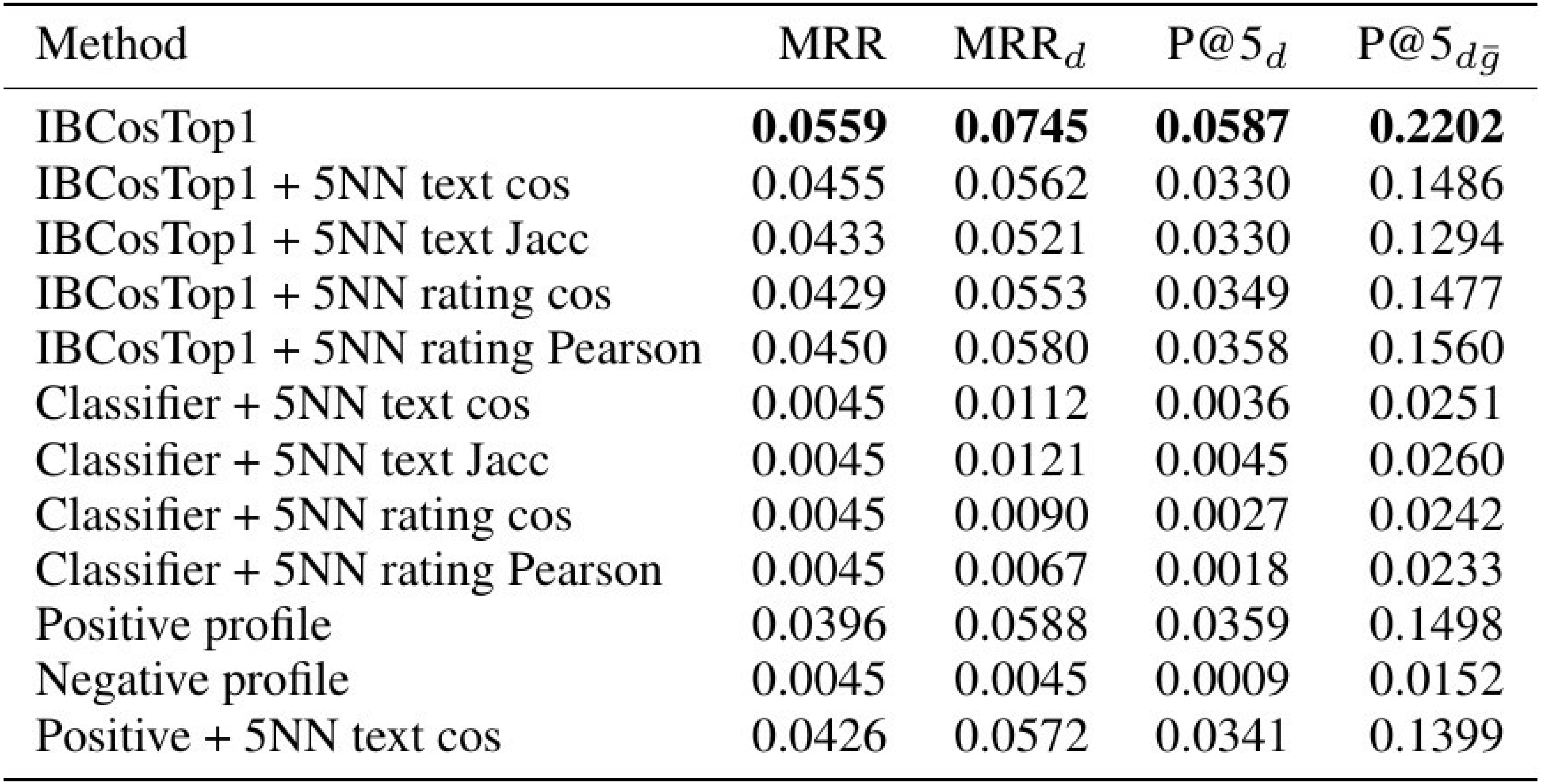




| Method | MRR | $\text{MRR}_d$ | $\text{P@5}_d$ | $\text{P@5}_{d\bar{g}}$ |
|---|---|---|---|---|
| IBCosTop1 | **0.0559** | 0.0745 | **0.0587** | **0.2202** |
| IBCosTop1 (given) | 0.0528 | **0.0955** | 0.0484 | 0.0780 |

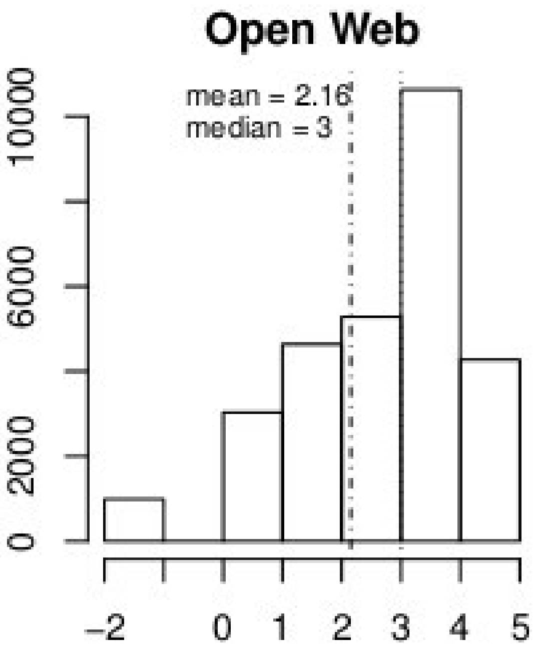
# Effect of ranking function

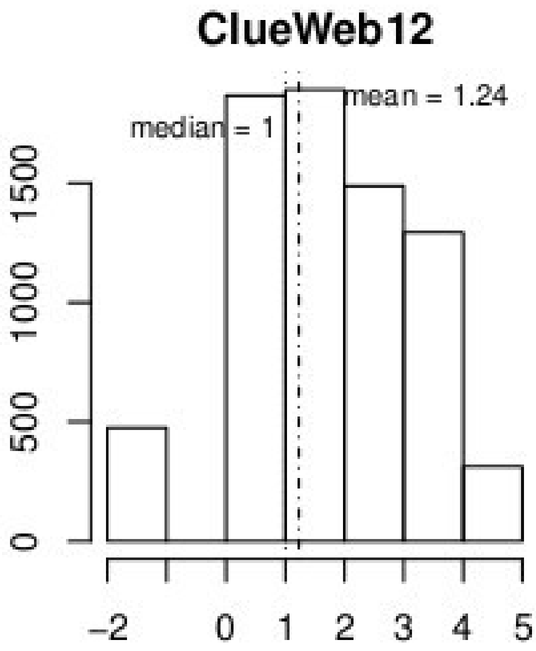
- (Low coverage of relevance assessment)
- 5-nearest neighbour outperform other k-neighbours
- Generating user profiles based on descriptions with negative rating gave the worst results

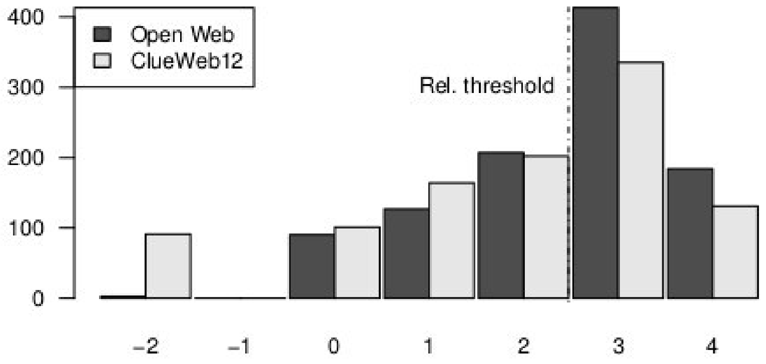


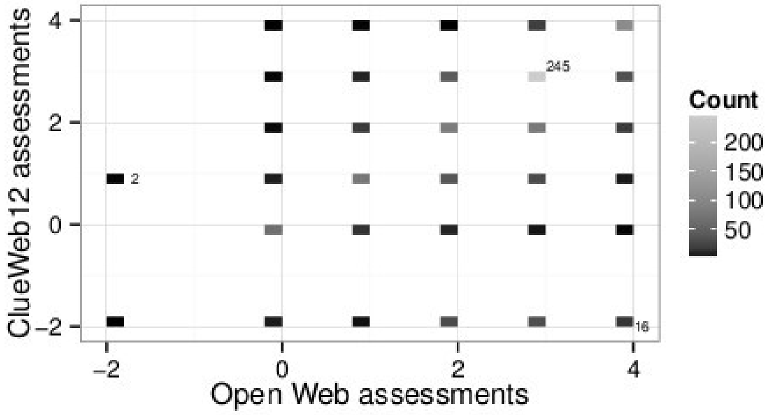
| Method | MRR | MRR$_d$ | P@5$_d$ | P@5$_{d\bar{g}}$ |
|---|---|---|---|---|
| IBCosTop1 | **0.0559** | **0.0745** | **0.0587** | **0.2202** |
| IBCosTop1 + 5NN text cos | 0.0455 | 0.0562 | 0.0330 | 0.1486 |
| IBCosTop1 + 5NN text Jacc | 0.0433 | 0.0521 | 0.0330 | 0.1294 |
| IBCosTop1 + 5NN rating cos | 0.0429 | 0.0553 | 0.0349 | 0.1477 |
| IBCosTop1 + 5NN rating Pearson | 0.0450 | 0.0580 | 0.0358 | 0.1560 |
| Classifier + 5NN text cos | 0.0045 | 0.0112 | 0.0036 | 0.0251 |
| Classifier + 5NN text Jacc | 0.0045 | 0.0121 | 0.0045 | 0.0260 |
| Classifier + 5NN rating cos | 0.0045 | 0.0090 | 0.0027 | 0.0242 |
| Classifier + 5NN rating Pearson | 0.0045 | 0.0067 | 0.0018 | 0.0233 |
| Positive profile | 0.0396 | 0.0588 | 0.0359 | 0.1498 |
| Negative profile | 0.0045 | 0.0045 | 0.0009 | 0.0152 |
| Positive + 5NN text cos | 0.0426 | 0.0572 | 0.0341 | 0.1399 |

# Archive Web vs Open Web evaluation









Thanks!