IMPROVING DIVERSITY RANKING USING SEVERAL RE-RANKINGS BASED ON SNIPPETS' SIMILARITIES (WORK IN PROGRESS)

ASHRAF BAH, BEN CARTERETTE
COMPUTER AND INFORMATION SCIENCES, UNIVERSITY OF DELAWARE

INTRODUCTION

Different users may type the same query but expect to satisfy different information need. In such situations, the retrieval system should not only retrieve and rank relevant documents, but also diversify the ranking. The goal of this work is to meet that diversity requirement by using snippets of the original documents, and applying several re-ranking techniques.

MATERIALS AND METHODS

For this preliminary phase, we use the TREC 2011 session track dataset. The data consists of **204 rankings** that we have to re-rank.
Our **proposed re-ranking methods** (each relies on cosine similarities):

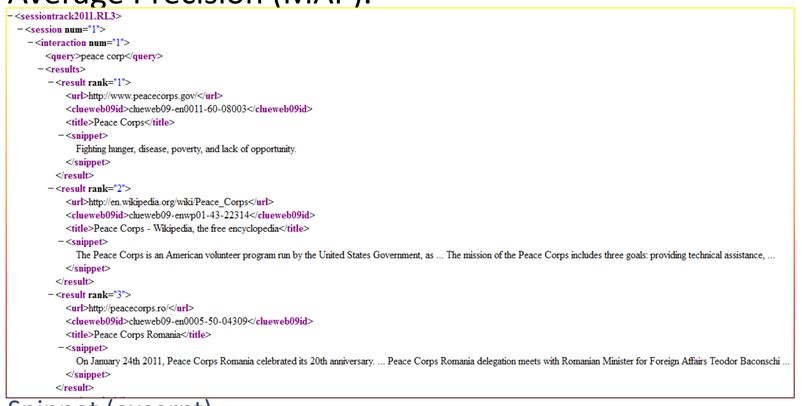
- Use Maximal Marginal Relevance – MMR (Carbonnel and Goldstein, 1998)

$$MMR \stackrel{\text{def}}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

- Use a variant of MMR
  Arg min$_i$ [avg$_j$ Sim(D$_i$,Dj)]
- Group the relevant documents in n groups. group1: the documents most similar to the top document;…; group n: the docs least similar to the top doc.
- First the top document. Second, the document least similar to the first. Third, the doc least similar to the third one, etc…

RESULTS

Currently, we are working on how to evaluate the results. We will be using measures including Estimated Reciprocal Rank (ERR), normalized Discounted Cumulative Gain (nDCG), and Mean Average Precision (MAP).



Snippet (excerpt)

```
METHOD 3 Re-ranking for ranking 0
Old rank -----> docId
0                    <clueweb09id>clueweb09-en0011-60-08003</clueweb09id>
2                    <clueweb09id>clueweb09-en0005-50-04309</clueweb09id>
1                    <clueweb09id>clueweb09-enwp01-43-22314</clueweb09id>
5                    <clueweb09id>clueweb09-en0074-24-11927</clueweb09id>
3                    <clueweb09id>clueweb09-en0005-88-05046</clueweb09id>
6                    <clueweb09id>clueweb09-en0009-98-35948</clueweb09id>
4                    <clueweb09id>clueweb09-en0028-43-18655</clueweb09id>
8                    <clueweb09id>clueweb09-en0012-95-03634</clueweb09id>
7                    <clueweb09id>clueweb09-en0002-55-22818</clueweb09id>
9                    <clueweb09id>clueweb09-en0133-67-12909</clueweb09id>


METHOD 3 Re-ranking for ranking 1
Old rank -----> docId
0                    <clueweb09id>clueweb09-en0011-60-08003</clueweb09id>
3                    <clueweb09id>clueweb09-en0024-37-21322</clueweb09id>
8                    <clueweb09id>clueweb09-en0109-28-37199</clueweb09id>
1                    <clueweb09id>clueweb09-en0001-91-23746</clueweb09id>
2                    <clueweb09id>clueweb09-en0058-45-05716</clueweb09id>
4                    <clueweb09id>clueweb09-en0018-64-08358</clueweb09id>
7                    <clueweb09id>clueweb09-en0024-90-40429</clueweb09id>
6                    <clueweb09id>clueweb09-en0005-88-05046</clueweb09id>
9                    <clueweb09id>clueweb09-en0028-43-18655</clueweb09id>
5                    <clueweb09id>clueweb09-enwp01-43-22314</clueweb09id>
```

Re-ranking Method 1 (excerpt)

```
METHOD 2 Re-ranking for ranking 0
Old rank -----> docId
0                    <clueweb09id>clueweb09-en0011-60-08003</clueweb09id>
2                    <clueweb09id>clueweb09-en0005-50-04309</clueweb09id>
1                    <clueweb09id>clueweb09-enwp01-43-22314</clueweb09id>
5                    <clueweb09id>clueweb09-en0074-24-11927</clueweb09id>
3                    <clueweb09id>clueweb09-en0005-88-05046</clueweb09id>
8                    <clueweb09id>clueweb09-en0012-95-03634</clueweb09id>
6                    <clueweb09id>clueweb09-en0009-98-35948</clueweb09id>
7                    <clueweb09id>clueweb09-en0002-55-22818</clueweb09id>
4                    <clueweb09id>clueweb09-en0028-43-18655</clueweb09id>
9                    <clueweb09id>clueweb09-en0133-67-12909</clueweb09id>


METHOD 2 Re-ranking for ranking 1
Old rank -----> docId
0                    <clueweb09id>clueweb09-en0011-60-08003</clueweb09id>
3                    <clueweb09id>clueweb09-en0024-37-21322</clueweb09id>
1                    <clueweb09id>clueweb09-en0001-91-23746</clueweb09id>
8                    <clueweb09id>clueweb09-en0109-28-37199</clueweb09id>
2                    <clueweb09id>clueweb09-en0058-45-05716</clueweb09id>
6                    <clueweb09id>clueweb09-en0005-88-05046</clueweb09id>
7                    <clueweb09id>clueweb09-en0024-90-40429</clueweb09id>
4                    <clueweb09id>clueweb09-en0018-64-08358</clueweb09id>
5                    <clueweb09id>clueweb09-enwp01-43-22314</clueweb09id>
9                    <clueweb09id>clueweb09-en0028-43-18655</clueweb09id>
```

Re-ranking Method 2 (Excerpt)

CONCLUSION

We have tried several re-ranking schemes using the similarities between the original documents' snippets. It would be interesting to see how well each method improves diversity. That evaluation is going to be our next step for the project.

LITERATURE CITED

- Carbonell, J. G., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In A. Mof-fat, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 335–336). Melbourne, Australia

CENTER FOR SCIENCE OF INFORMATION
A National Science Foundation Science and Technology Center
soihub.org