Fight Against COVID-19 Misinformation via Clustering-Based Subset Selection Fusion Methods

Yidong Huang[a],Qiuyu Xu[a], Shengli Wu[a], Christopher Nugent[b] and Adrian Moore[b]

[a]School of Computer Science, Jiangsu University, China
[b]School of Computing, Ulster University, UK

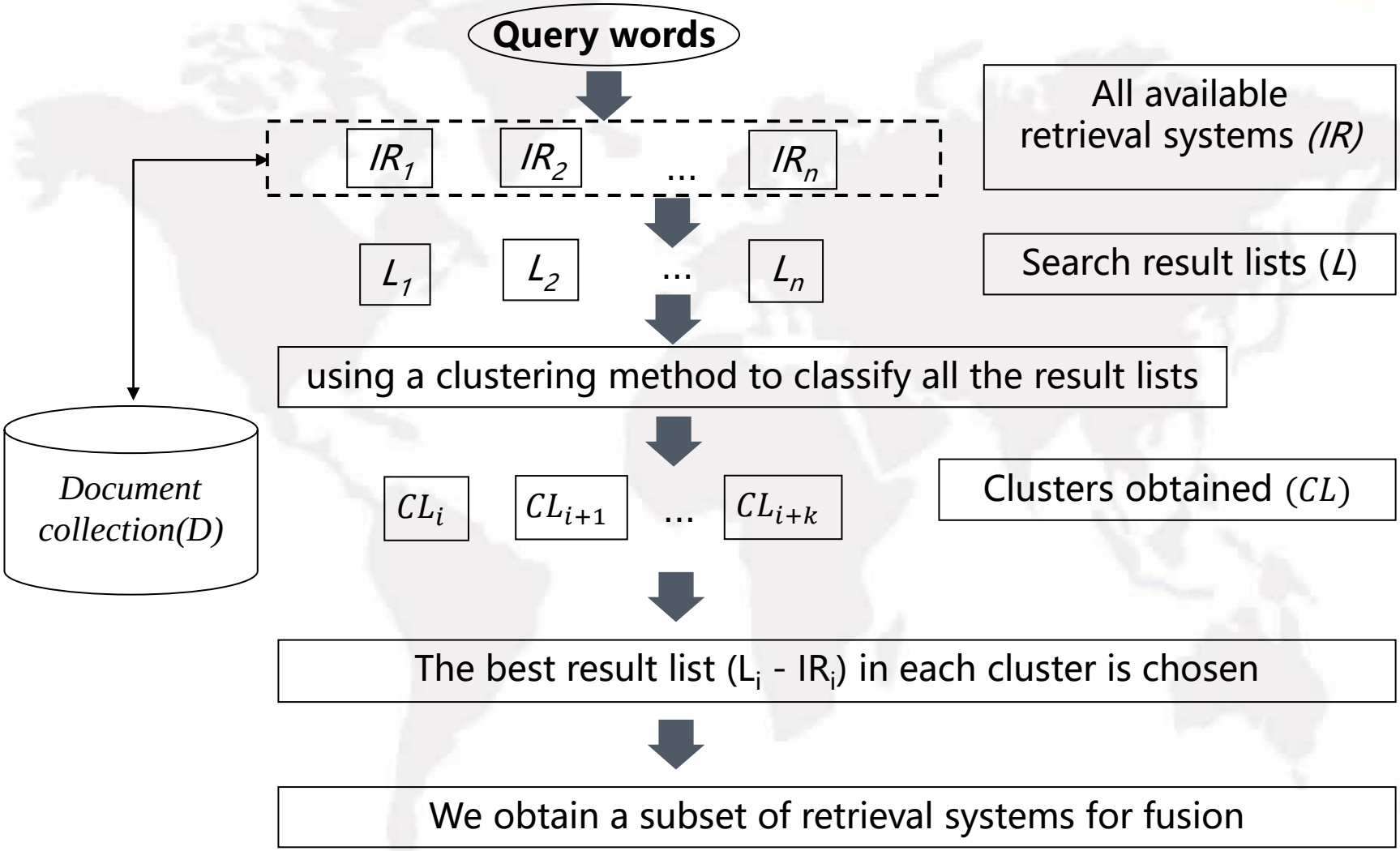
Background

- The worldwide COVID-19 pandemic has brought about a lot of changes in people's life. It also emerges as a new challenge to information search services.
- Responsibility of search engines is crucial because many people make decisions based on the information available to them.
- Therefore, we explore fusion-based approaches for medical retrieval tasks.

Proposed method

- Fusion is an effective technique for retrieval performance improvement.
- It costs more than a typical retrieval system because multiple component retrieval systems are involved.
- A clustering-based approach is proposed for selecting a subset of retrieval systems from all available ones.

Workflow of the research work



Classify systems/results and select

- First all component systems/results are set into clusters by considering their similarity.
- The second step is to choose a group of retrieval systems for fusion. In this step, we can take top performers from different clusters, thus both performance of component systems (good performers in a cluster) and diversity in the selected systems (chosen from different clusters) can be considered in tandem.

The clustering-based model

For our investigation, K-means is a good option for clustering relatively a small number of retrieval systems (e.g., the data set of Health Misinformation Track in TREC 2000 comprises 51 runs) and the Euclidean distance between them is well-defined for clustering.

Euclidean distance of two system/result

- Calculate the distance between two systems/results by

$$Dist(L_1, L_2) = \sum_{i=1}^{|D|} \sqrt{(s_1(d_i) - s_2(d_i))^2}$$

C1 and C2

C1: Select the best performer L (in MAP, or Mean Average Precision) in each cluster, discard the three worst performers.

C2:An improved variant of K-means is used for clustering [39]. Exact number of clusters generated to match the number of retrieval systems for selection.

Data and Experimental setting

- In November 2020, TREC held a Health Information Track. It used the documents found in the CommonCrawl News crawl from January 1, 2020 to April 30, 2020. The crawl contains news articles from web sites all over the world.
- There are 50 queries. All include number, title, description, answer, evidence, and narrative.
- Apart from C1 and C2, two baseline methods, Top_J and Top_MAP, are also tested. Both select candidates based on performance (different measures used for evalaution)

Experimental results

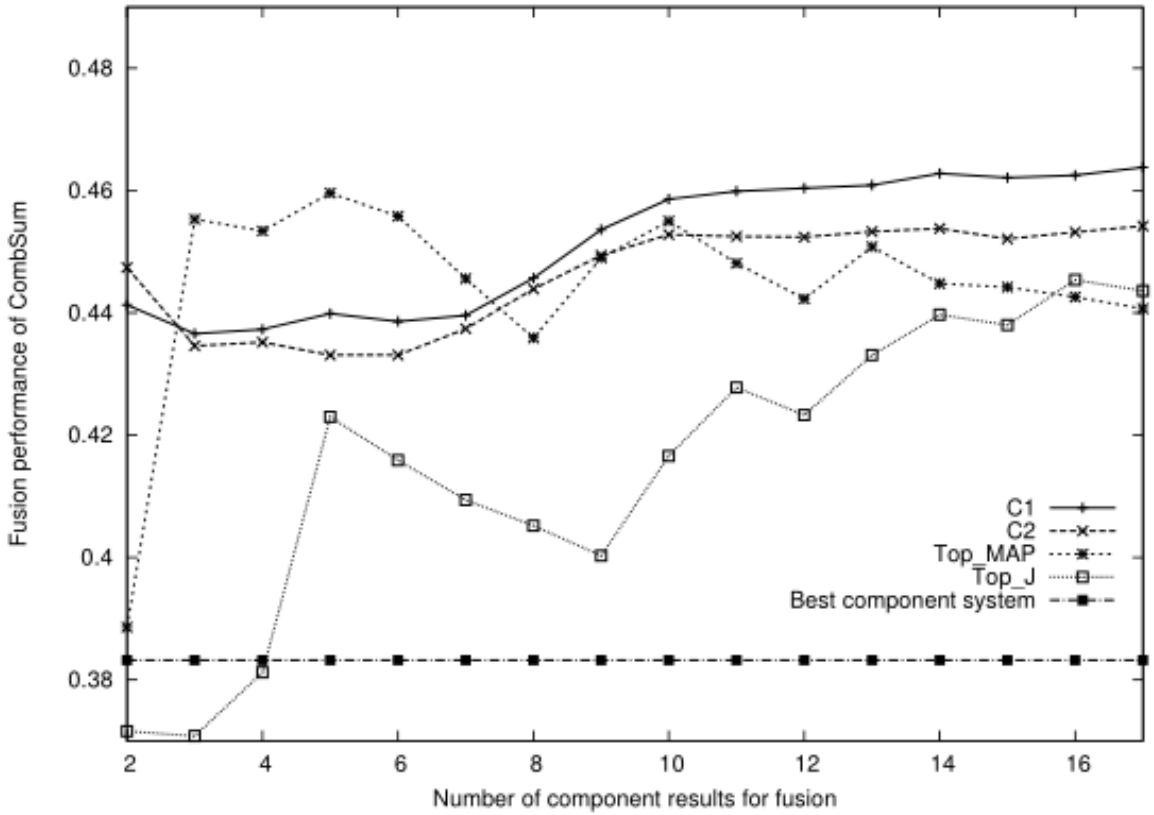


Figure 2: Comparison of four subset selection methods (component results are fused by CombSum and fusion results are evaluated by MAP)

Experimental results

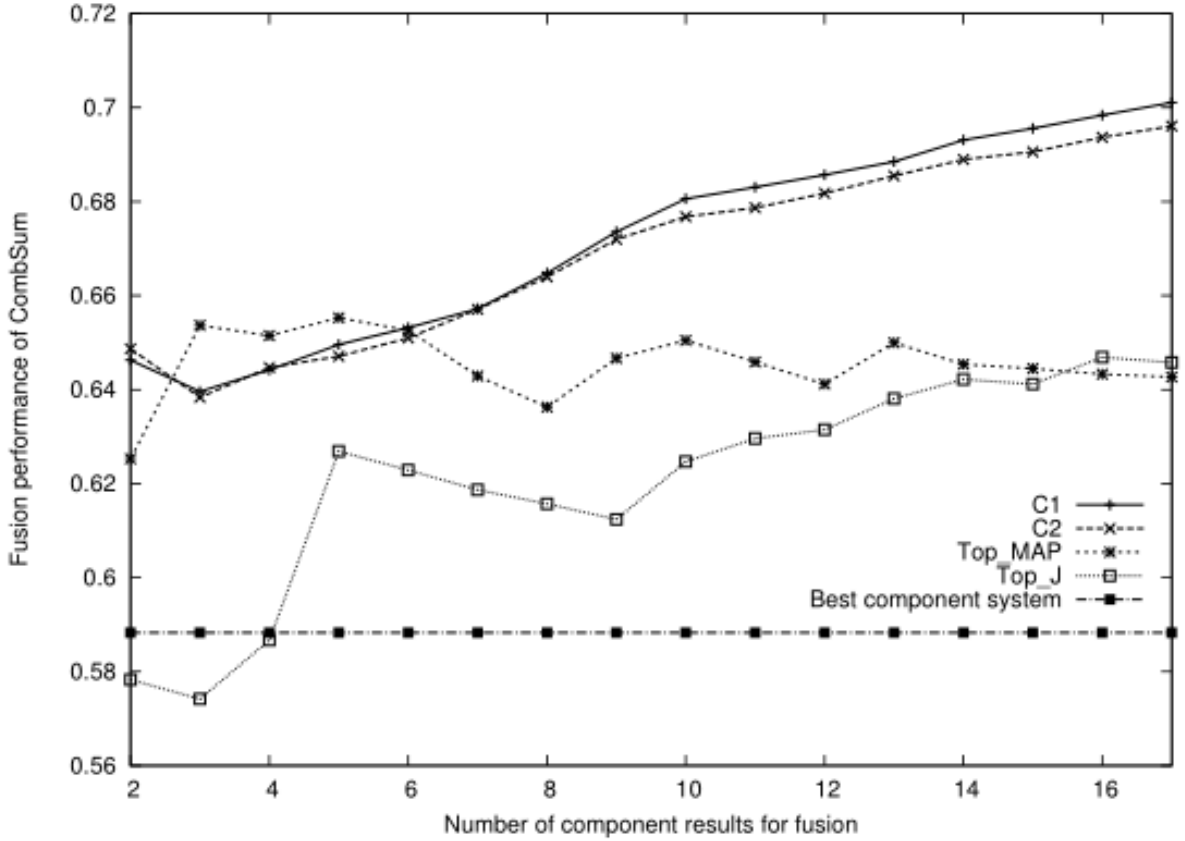


Figure 3: Comparison of four subset selection methods (component results are fused by CombSum and fusion results are evaluated by CAM)

Experimental results

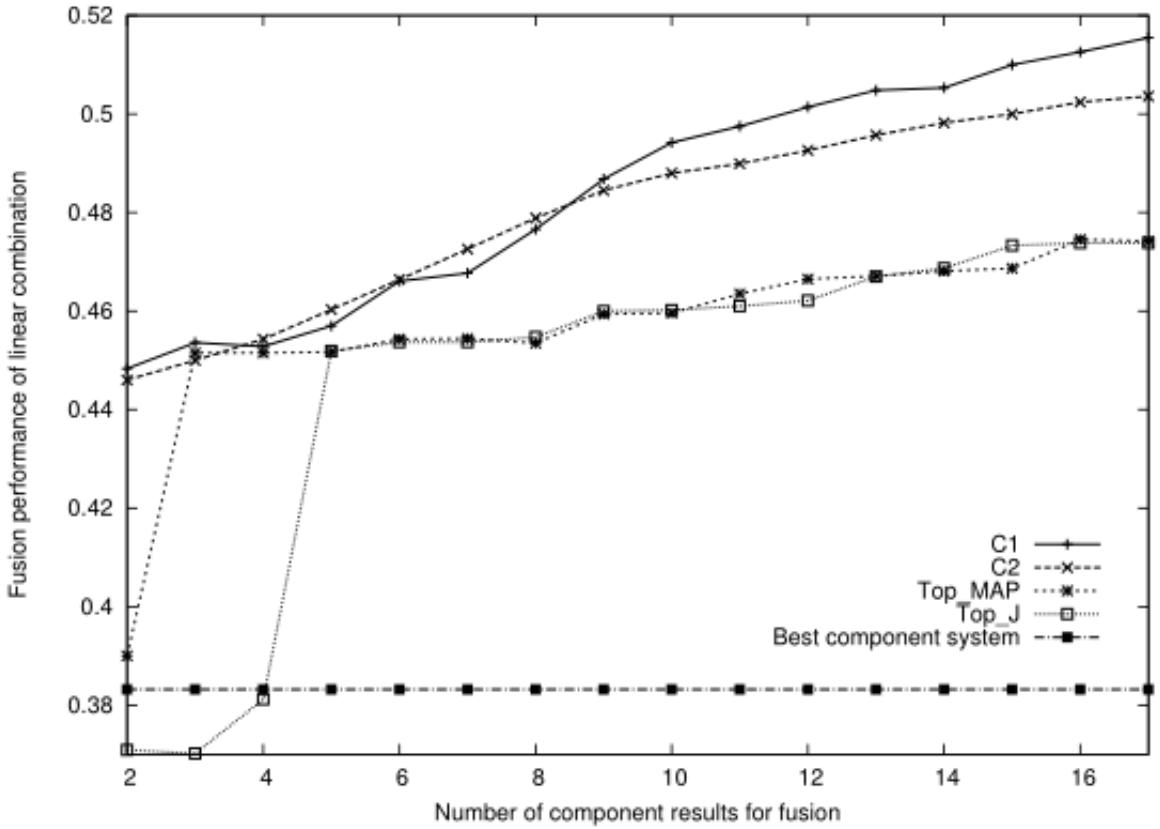


Figure 4: Comparison of four subset selection methods (component results are fused by linear combination and fusion results are evaluated by MAP)

Experimental results

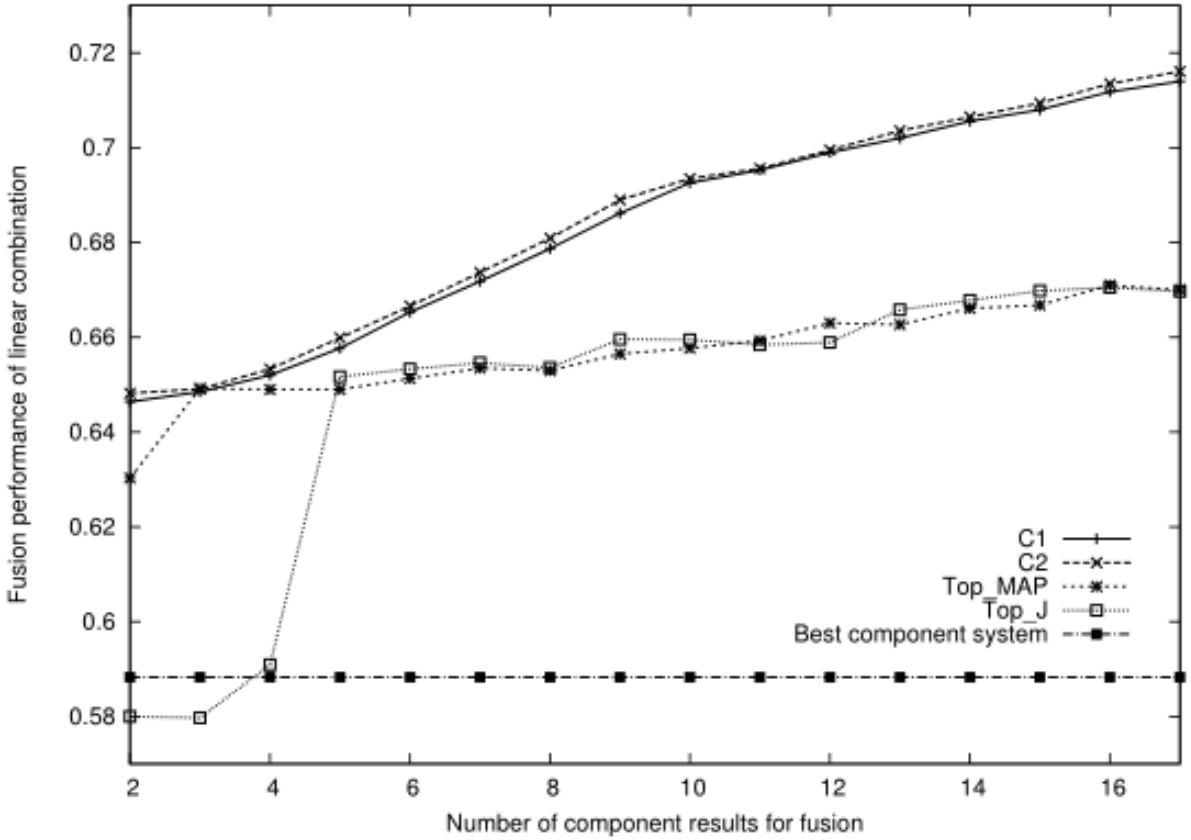


Figure 5: Comparison of four subset selection methods (component results are fused by linear combination and fusion results are evaluated by CAM)

表3. 2 实验所用文档集的详细信息

Experimental results

Table 2

Pairwise comparison of subset section methods(A figure in bold indicates that the difference between the two methods is significant at the .05 level; T_M denotes Top_MAP; T_J denotes Top_J)

Method/Measure	C1:C2	C1:T_M	C1:T_J	C2:T_M	C2:T_J	T_M:T_J
CombSum/MAP	**1.20%**	1.57%	**8.72%**	0.36%	**7.42%**	**3.19%**
CombSum/CAM	**0.37%**	**4.14%**	**8.24%**	**3.75%**	**7.83%**	**3.95%**
LN/MAP	**-0.23%**	**4.88%**	**7.42%**	**5.12%**	**7.67%**	0.15%
LN/CAM	**-0.22%**	**4.05%**	**5.71%**	**4.28%**	**5.94%**	1.59%

表3. 2 实验所用文档集的详细信息

Experimental results

Table 3

Analysis of four subset selection methods (each triplet includes MAP/average pairwise distance values of component result lists/Combi)

Number	C1 MAP/Dist/Combi	C2 MAP/Dist/Combi	Top_MAP MAP/Dist/Combi	Top_J MAP/Dist/Combi
2	0.374/3.358/0.859	0.374/3.742/0.900	0.377/1.186/0.628	0.360/1.206/0.608
3	0.358/3.146/0.844	0.354/3.182/0.813	0.373/2.732/0.790	0.361/0.919/0.578
4	0.340/3.168/0,793	0.333/3.254/0.793	0.370/2.628/0.775	0.347/1.831/0.658
5	0.318/3.361/0.785	0.311/3.425/0.782	0.368/2.439/0.752	0.354/2.453/0.735
6	0.296/3.576/0.779	0.290/3.621/0.776	0.366/2.449/0.750	0.355/2.218/0.710
7	0.278/3.764/0.775	0.273/3.823/0.775	0.365/2.361/0.739	0.355/2.018/0.689
8	0.262/3.959/0.775	0.257/4.002/0.773	0.364/2.242/0.725	0.354/1.892/0.674
9	0.248/4.126/0.775	0.243/4.152/0.771	0.363/2.358/0.736	0.352/1.806/0.662
10	0.235/4.244/0.770	0.230/4.152/0.754	0.361/2.376/0.735	0.353/2.077/0.693
11	0.224/4.335/0.765	0.218/4.360/0.760	0.360/2.340/0.730	0.355/2.244/0.713
12	0.213/4.409/0.759	0.208/4.428/0.754	0.359/2.303/0.725	0.351/2.189/0.702
13	0.203/4.469/0.752	0.199/4.488/0.749	0.356/2.369/0.728	0.351/2.316/0.716
14	0.194/4.425/0.735	0.189/4.536/0.741	0.353/2.348/0.722	0.351/2.380/0.723
15	0.185/4.577/0.740	0.180/4.584/0.734	0.351/2.401/0.725	0.349/2.400/0.722
16	0.176/4.629/0.733	0.181/4.503/0.726	0.348/2.422/0.723	0.347/2.458/0.726
17	0.187/4.521/0.736	0.172/4.574/0.722	0.346/2.447/0.723	0.345/2.490/0.727

Conclusions

- In this paper, we have presented clustering-based methods for selecting a subset of component retrieval systems from all available ones to achieve good fusion performance.
- One major characteristic of the proposed methods is they take both performance of component systems and dissimilarity among them into consideration at the same time. It is better than the approaches that considers performance only.
- It also demonstrates that data fusion is a good approach for this Health Misinformation task.

Future work

- One is to investigate the relationship between component system performance and dissimilarity among component results. If a more precise relationship can be set up for them, then it is possible to find more efficient and effective system selection methods for fusion.
- Another direction is to design an unsupervised version of such methods. At present, generating a usable training dataset can be very costly because relevance judgment by human referees is required for those retrieved documents.

Thank you!