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

Yidong Huang[1], Qiuyu Xu[1], Shengli Wu[1], Christopher Nugent[2] and Adrian Moore[2]

[1]School of Computer Science, Jiangsu University, China
[2]School of Computing, Ulster University, UK

Abstract

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. This is because up to now our understanding about the virus is still limited, and there is a lot of misinformation online. In such a situation, how to provide useful and correct information to the public is not straightforward. Responsibility of search engines is crucial because many people make decisions based on the information available to them. In this piece of work, we try to improve retrieval quality via the data fusion technique. Especially, a clustering-based approach is proposed for selecting a subset of systems from all available ones for finding relevant, credible, and correct documents. Experimented with a group of runs submitted to the 2020 TREC Health Misinformation Track, we demonstrate that data fusion is a very beneficial approach for this task, whether measured by some traditional metrics such as MAP or some task specific metrics such as CAM. When choosing 17 runs, which is one third of all component retrieval systems available, the linear combination method is better than the best component retrieval system by 31.42% in MAP and 21.72% in CAM. The proposed methods are also better than the state-of-the-art subset selection method by a clear margin.

Keywords

Data Fusion, Information Retrieval, Health Misinformation, Credibility, COVID-19

1. Introduction

Since the initial cases were discovered at the end of 2019, within two years COVID-19 has been spreading globally to almost all major countries and territories, with over two hundred million confirmed cases and over four million deaths so far. Such a unprecedented pandemic has impacted people's life significantly. For many, it is very valuable to get useful and correct information about the virus. However, this may not be as straightforward as it looks, because there are still a lot of things we do not know about the virus and considerable misinformation exists on the web [1] and disseminates on social media [2, 3]. The consequences of such infodemic is very harmful to the society and has negative impact on our fight against the pandemic. Therefore, it is necessary to understand this phenomenon and develop some measures to fight against it.

Some research on this issue has been conducted so far. A few of them focus on observation and analysis while some others focus on misinformation detection. For example, [4] analysed a Singapore-based COVID-19 Telegram group with more than 10,000 participants. There are

ROMCIR 2022: The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2022: the 44th European Conference on Information Retrieval, April 10-14, 2022, Stavanger, Norway

a few observations and one of them is that authority-identified misinformation is rare. Both [5] and [6] analysed misinformation on Chinese Sina Weibo, while [7] did it on Twitter. In [8], machine learning techniques including decision tree and convolutional neural network based models were used to classify COVID-19 related information and misinformation.

In 2020, TREC (Text REtrieval Conference) [1] held two COVID-19 related tracks: COVID [9] and Health Misinformation [10]. In this piece of work, we focus on the ad-hoc retrieval task in the Health Misinformation Track. 51 runs were submitted to this task by eight research groups. 50 queries were used for this task. Two baseline runs, BM25 desc and BM25 title, were submitted by the UWaterlooMDS group on behalf of the track organizers. Those submitted runs are available on TREC's web site. They provide us a very good opportunity to investigate how system-level data fusion can improve retrieval quality in this task.

Data fusion has been widely used in information retrieval for different tasks [11, 12, 13, 14]. Fusion performance is affected by many factors including fusion methods, each of the component retrieval systems (results) involved, the number of component systems in total, evaluation metrics, among others. In this study, we investigate how to improve retrieval performance in this task by using the data fusion technology. More specifically, our research question is: given a large collection of retrieval systems, how can we choose a subset of them for effective and efficient fusion? This has rarely been investigated before. To our knowledge, [15] is the only one that addressed this issue. Because there are many different retrieval models, many components such as name recognition, phrases, semantic relations of concepts, different techniques for document credibility, and many others, it is possible to build/collect relatively a large number of component retrieval systems for fusion. However, the efficiency of a fusion-based system decreases when more component retrieval systems are involved. For such a retrieval system, both performance and efficiency need to be considered. It is an important problem that deserves research. In this paper, we propose a clustering-based method to deal with this problem. First we apply K-means to divide all the systems into a given number of clusters, then one representative is chosen from each cluster to form a group for fusion. In this way, both system performance and diversity among systems can be considered at the same time. It is able to obtain better fusion performance than those selection methods that only consider system performance only as in [15]. Experimented with all 51 runs submitted to the 2020 TREC Health Misinformation Track, our results show that the method is very effective.

The rest of this paper is organized as follows: related work is discussed in Section 2. The proposed clustering-based data fusion method is detailed in Section 3. Section 4 presents experimental settings and results of the proposed method and some other baseline methods. Section 5 presents some more analytical results on the clustering method and the clusters generated on the 2020 TREC Health Misinformation data set. Section 6 concludes the paper.

2. Related Work

In this paper, we investigate how to apply data fusion to COVID-19 related health information retrieval with considerable misinformation inside the collection. Therefore, we review some previous work on COVID-19 misinformation detection and credible information retrieval. After

[1] Its web site is located at https://trec.nist.gov/

that, we review some data fusion methods and some of its application in medical information retrieval.

2.1. Misinformation Detection & Credible Information Retrieval

Since confirmed COVID-19 cases first occurred at the end of 2019 and began to spread around the world afterwards, a lot of rumours, misinformation, and disinformation turn up on social media and the Web, and circulate in certain communities. How to detect misinformation becomes a key issue in medical information retrieval. Various machine learning techniques have been used to detect misinformation. In [8], both decision tree classifiers and convolutional neural networks were used to classify COVID-19 related information and misinformation. [16] applied the Elaboration Likelihood Model with four types of features: linguistic, topical, sentimental, and behavioural features. It was found that behavioural features are more informative than linguistic features for their detection. [17] proposed a deep learning network that could leverage both visual and textual information. In their semantic and task level attention model, three branches were defined to extract features of different types. An ensemble method was also used for the detection. Some more work were presented in [18, 19] among others.

Because the Web is an open environment, documents on the Web may be in a variety of quality. Web documents' credibility has been a research issue for the last two decades [20]. In this article, we employ the term credibility with the meaning it has in [20], where it is described as a general concept that encompasses trustworthiness, expertise, quality, and reliability. Such a term has been adopted in computer science for many years [21, 22, 23, 24], and it has special importance in the Information Retrieval/Web search community [20].

For medical retrieval systems, document credibility is also an important and challenging issue [25]. To retrieve documents that are both relevant and credible, usually a two-stage process is taken. First documents are retrieved by only considering their relevance to the query. Then the documents are re-ranked by considering both relevance and credibility. Some traditional models such as BM25 can be used for relevance-concerned retrieval, while credibility of documents can be predicted by some machine learning methods [26, 27, 28].

2.2. Data Fusion

Data fusion methods can be divided into two categories: supervised and unsupervised methods. CombSum [29], CombMNZ [29], and the Reciprocal Rank [30] are typical unsupervised methods, while linear combination [31] is a typical supervised method. Unsupervised methods are easy to use, while supervised methods are suitable for various situations in which unsupervised methods do not perform well.

Data fusion methods have been applied to various tasks in information retrieval [11, 12, 13, 14]. It is also popular for medical retrieval tasks [32, 33, 34]. Some form of data fusion techniques are also used in those runs submitted to the 2020 TREC Health Misinformation Track, which we use for the experiments. For example, both runs, CiTIUSCrdRelAdh and CiTIUSSimRelAdh, submitted by the CiTIUS group [35], used Borda Count, to combine two types of rankings: relevance and reliability (credibility & correctness). For the h2oloo group [36], query expansion and two types of machine learning technologies were used for re-ranking. All eight runs

submitted were various combinations of them and the BM25 baseline run, in which equal or simple inequal weights were used. Similar situation exists in some other submissions.

Usually, the number of component retrieval systems involved is a good indicator of the complexity of a fusion-based system. With equal final performance, it is preferable to have fewer component retrieval systems involved. [15] investigated how to choose a subset from a large group of retrieval systems for better fusion performance, although those retrieval systems/results are not for medical retrieval tasks. A DCG-like (Discounted Cumulative Gain, a commonly used metric in information retrieval evaluation) measure was defined for the selection purpose. One limitation of this research is: it only considered performance of those candidate systems, but not diversity of those systems (results) chosen. As a matter of fact, both component system performance and dissimilarity among component systems (results) affect fusion performance significantly.

In this piece of work, we investigate how to achieve the best possible results by using the data fusion technology for this misinformation retrieval task. We focus on the problem of subset selection for effective fusion. The task is: for a group of N retrieval systems, how to select n ($n < N$) of them to obtain the best fusion performance? This task is the same as that in [15]. However, we propose a clustering-based method for this task, which is different from [15]. Both theoretical analysis and empirical investigation demonstrate that our proposed method is more effective than the one proposed in [15]. Besides, [15] used four data sets from CLEF (Cross-Language Evaluation Forum) [2] for their empirical investigation.

This piece of work is also different from those applied data fusion methods for medical retrieval [11, 12, 13, 14, 35, 36]. All of them empirically investigated the effectiveness of a few typical data fusion methods for the chosen task. Choosing a subset from a large group of candidate systems is not a research task in those studies.

3. Subset Selection for Fusion

For a group of information retrieval systems, how to select a subset for best possible fusion effectiveness is a challenging task. For example, if we have 50 retrieval systems and try to select 10 of them for better fusion performance, then the number of possible combinations is huge. As a matter of fact, the exact number to this question is 50*49*...*41, or 37,276,043,023,296,000. Therefore, it may not be possible to test all of them. Instead of doing an exhaustive search to try to find the best possible solution, to develop and use some heuristic methods is more realistic.

3.1. Top_J

In this vein, [15] defined a DCG-like measure, which is referred to as J-measure later in this paper. It is defined as

$$J(L) = \sum_{i=1}^{|L|} (1 - \frac{ln(i)}{ln|L|}) * rel(d_i) \quad (1)$$

[2] Its web site is located at https://www.clef-campaign.org/

where L is a ranked list of documents for a given query, $|L|$ is the number of documents in L, $d_1, d_2, ..., d_{|L|}$ are documents in L, and $rel(d_i)$=1 if d_i is relevant to the query and $rel(d_i)$=0 otherwise. For a group of resulting lists, J values can be used to evaluate and select component results (and corresponding retrieval systems) for fusion. This selection method is referred to as Top_J in this paper. Top_J is reasonable because it is found that better component systems/results usually lead to better fusion performance [37, 38].

3.2. Clustering-Based Subset Selection for Fusion

Previous research [38] found that performance of component systems/results is not the only factor that affects fusion performance. Diversity of the component retrieval systems/results is also a factor that affects fusion performance significantly, but it is not considered in Top_J. To incorporate diversity to the selection process, we propose clustering-based methods. There are two major steps involved. First all component systems/results are set into clusters by considering their similarity. Consequently, we can expect that the systems/results in the same cluster are similar and the systems/results not in the same cluster are very different. 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.

Now let us see how to perform the clustering method for those retrieval systems. We assume that the properties of a retrieval system is fully reflected by the results it retrieves. For two retrieval systems, we can observe the similarity/dissimilarity of the two ranked lists of results they generate for the same query. Scoring is used in this work and we can define the Euclidean distance to measure the dissimilarity of two resulting lists.

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

where L_1 and L_2 are retrieved result lists from two retrieval systems for the same collection D and same query q, $|D|$ is the number of documents in D, $s_1(d_i)$ is the score that d_i obtains in L_1, and $s_2(d_i)$ is the score that d_i obtains in L_2. For all the documents in D that do not appear in L_1 (or L_2), we need to define a default score (e.g., zero) for them. $Dist(L_1, L_2)$ denotes the distance between L_1 and L_2, which is a good indicator of the dissimilarity between L_1 and L_2. Although not used here, ranking information is an alternative for the same purpose.

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. Most clustering methods such as K-means requires a pre-defined value as the number of clusters. That value needs to be considered carefully. When the number of clusters are very small, it is possible that quite different results have to go to the same cluster. Therefore, such a situation should be avoided even we just need a small number of component results for fusion. On the other hand, if too many clusters are generated, then each cluster will become very small. Considering that there are 51 runs in the data set used for the experiment, we decide to generate 17 clusters.

Table 1
Statistics of the data set (all 51 runs submitted to the adhoc task of the Health Information Track in TREC 2020)

Measure	Best Run	AVE.	STDV
MAP	0.3832 (h2oloo.m5)	0.2118	0.1199
CAM	0.5883 (h2oloo.m5)	0.3606	0.1399

Thus each cluster has three result lists on average. It would give us some flexibility for the selection of candidates. For the time being, we take a simple selection method: first we select the best performer L (in MAP, or Mean Average Precision) in all the clusters. Then we removed the cluster to which L belongs. For the remaining clusters repeat the above process until we get enough result lists. In this way both performance of component result lists and their diversity can be considered at the same time. This method is referred to as C1 later in this paper.

The quality of clusters generated by K-means is determined by the initial K points, which are chosen randomly. In order to improve the quality of clustering, we use a variant of K-means presented in [39]. Its main idea is to generate J solutions by K-means. Then the best is chosen from those J candidates. It is a little more complicated than standard K-means but usually produce clusters in better quality. It is referred to as C2 later in this paper.

4. Experimental Settings and Results

In this section we present the setting and results of the experiment carried out to validate the proposed methods. Especially, the data set used is the ad-hoc task of the Health Misinformation Track in TREC 2020, we would demonstrate the applicability of the proposed methods to this special information seeking task.

4.1. Experimental Settings

In November 2020, TREC held a Health Information Track [10]. The track 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.

The topics (queries) for this track focused on the consumer health search domain relevant to COVID-19. Fifty topics with a fixed structure were provided. All include number, title, description, answer, evidence, and narrative. Fig. 1 gives an example of the topic used. The title field has the form of a pair of treatment and disease. The description is formulated as a question, which contains treatment, effect, and disease. The answer corresponds to the medical consensus at the time of topic creation. Finally, the remaining fields were not intended to be used by the retrieval systems, but only by human assessors to produce relevance judgment document "qrels".

It set two tasks: total recall and ad-hoc retrieval. In this study, we use all 51 runs submitted to the ad-hoc retrieval task by eight research groups. Their statistics are summarized in Table 1.

Apart from C1 and C2, two baseline methods, Top_J and Top_MAP, are also tested. The

```
<topic>
 <number>0</number>
 <title>Ibuprofen COVID-19</title>
 <description>Can ibuprofen worsen COVID-19?</description>
 <answer>no</answer>
 <narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever
  and treat pain or inflammation. Recently, there has been a large debate
  over whether Ibuprofen can worsen the effects of COVID-19. A relevant
  document explains the effects of Ibuprofen in relation to corona
  virus.</narrative>
</topic>
```

Figure 1: Example of topic for the Health Misinformation Track 2020

common ground of Top_J and Top_MAP is that both of them only consider performance of component systems but not diversity of the selected systems. However, slightly different from Top_J, Top_MAP chooses retrieval systems based on their MAP values.

Two measures, MAP (Mean Average Precision) and CAM (Convex Aggregation Measure), are used for retrieval results evaluation. MAP is a classical measure commonly used for effectiveness evaluation of retrieval results, while CAM considers multiple aspects of a retrieved result list [40]. It is defined as

$$CAM(L) = M_{rel}(L)/3 + M_{cor}(L)/3 + M_{cre}(L)/3 \tag{3}$$

where L is a ranked list of documents with multi-aspect labels, M_{rel}, M_{cor}, and M_{cre} denote respectively any valid relevance, correctness, and credibility evaluation measures. In this study, we follow the instantiation of TREC by using nDCG for each individual aspect. That is to calculate M_{rel} as standard nDCG with respect to relevance, M_{cor} as standard nDCG with respect to correctness labels, and M_{rel} as standard nDCG with respect to credibility.

4.2. Experimental Methodology and Results

When the resulting lists are chosen from those clusters, we use CombSum, CombMNZ, and linear combination to fuse them.

For the same document collection D and a group of retrieval systems ir_i for ($1 \leq i \leq n$). All retrieval systems ir_i ($1 \leq i \leq n$) search D for a given query q and each of them provides a ranked list of documents $L_i = < d_{i1}, d_{i2}, ..., d_{im} >$. Assume that a relevance score $s_i(d_{ij})$ is associated with each of the retrieved documents in the list. CombSum [29, 41] uses the following equation

$$g(d) = \sum_{i=1}^{n} s_i(d) \tag{4}$$

to calculate scores for every document d. Here $s_i(d)$ is the score that ir_i assigns to d. If d does not appear in any L_i, then a default score (e.g., 0) must be assigned to it. After that, every document d obtains a global score $g(d)$ and all the documents can be ranked according to the global scores they obtain.

CombMNZ [29, 41] uses the equation

$$g(d) = m * \sum_{i=1}^{n} s_i(d) \tag{5}$$

to calculate scores. Here m is the number of results in which document d appears.

The linear combination method [31] uses the equation below

$$g(d) = \sum_{i=1}^{n} w_i * s(d) \tag{6}$$

to calculate scores. w_i is the weight assigned to system ir_i. Obviously, the linear combination is a general form of CombSum. If all the weights w_i are equals to 1, then the linear combination is the same as CombSum. Note that how to assign weights to different retrieval systems is an important issue. We use multiple linear regression to train weights [31] for it.

Let a training data set comprises a collection of l documents (D), a group of m queries (Q), and a group of n information retrieval systems (IR). For each query q^i, all information retrieval systems ir_j $(1 \leq j \leq n)$ provide their estimated relevance scores to all the documents in the collection. Therefore, we have (s^i_{1k}, s^i_{2k},..., s^i_{nk}, y^i_k) for i = (1, 2, ..., m), k = (1, 2, ..., l). Here s^i_{jk} stands for the score assigned by retrieval system ir_j to document d_k for query q^i; y^i_k is the judged relevance score of d_k for query q^i. If binary relevance judgment is used, then it is 1 for relevant documents and 0 otherwise.

$Y = \{y^i_k; i = (1, 2, ..., m), k = (1, 2, ..., l)\}$ can be estimated by a linear combination of scores from all component systems. Consider the following quantity

$$\mathcal{F} = \sum_{i=1}^{m} \sum_{k=1}^{l} [y^i_k - (\hat{\beta}_0 + \hat{\beta}_1 s^i_{1k} + \hat{\beta}_2 s^i_{2k} + ... + \hat{\beta}_n s^i_{nk})]^2$$

when $\mathcal{F}$ reaches its minimum, the estimation is the most accurate. β_0, β_1, β_2,..., and β_n, the multiple linear regression coefficients, are numerical constants that can be determined from observed data.

In the least squares sense the coefficients obtained by multiple linear regression can bring us the optimum fusion results by the linear combination method, since they can be used to make the most accurate estimation of the relevance scores of all the documents to all the queries as a whole [31]. β_j can be used as weights for retrieval systems ir_j $(1 \leq j \leq n)$ for fusion.

Score normalization is a necessary step for fusing all the result lists. For any of the component result lists, the retrieved documents are assigned scores using 1/(rank(d)+60), where rank(d) is the ranking position of document d. It is proposed in [30].

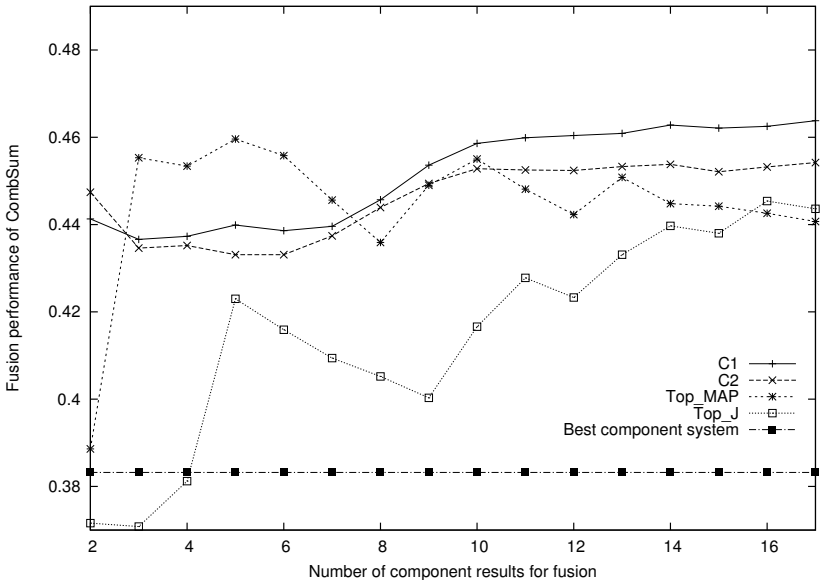

Figure 2: Comparison of four subset selection methods (component results are fused by CombSum and fusion results are evaluated by MAP)

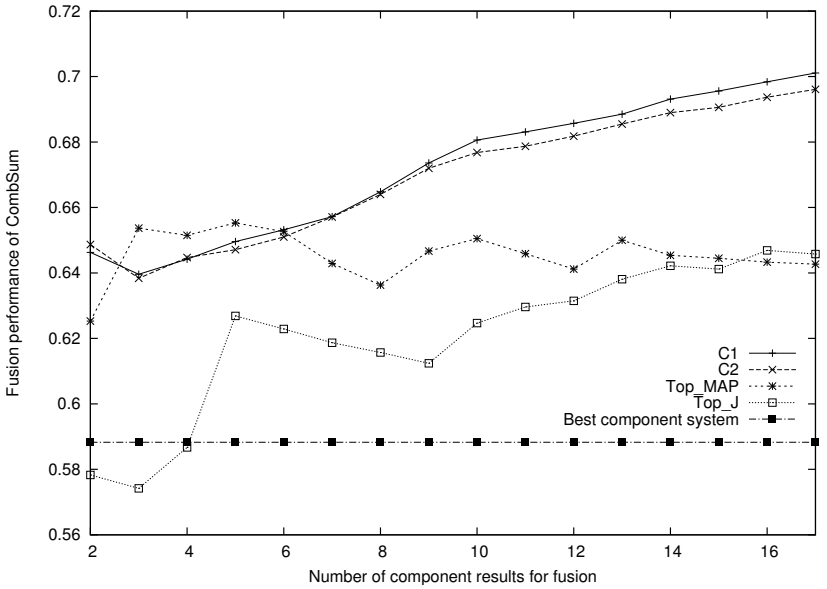

Figure 3: Comparison of four subset selection methods (component results are fused by CombSum and fusion results are evaluated by CAM)

All the queries are divided into two groups: odd-numbered and even-numbered, then two-fold cross-validation [3] is applied. There are uncertainty involved in C1 and C2. We run both of them

[3]N-fold cross-validation is a commonly used methodology in machine learning for training and testing a model on the same dataset.

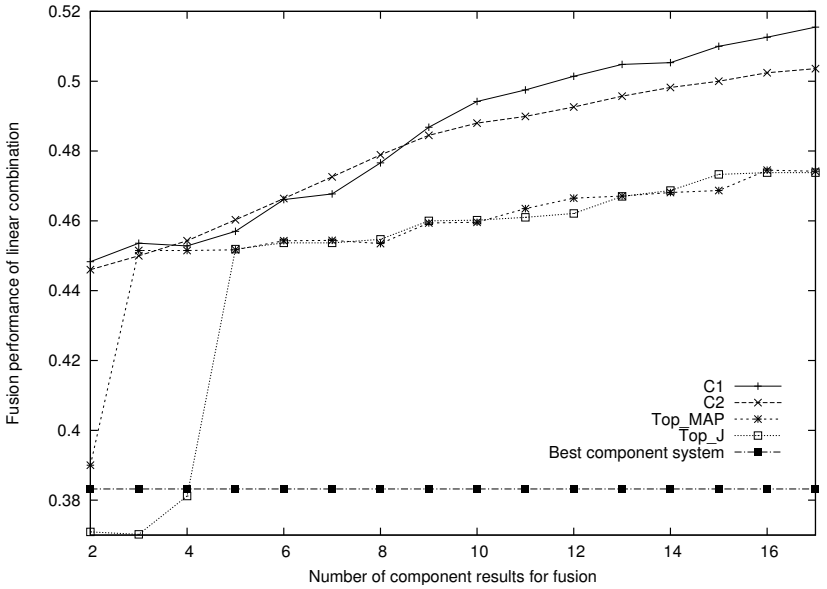

Figure 4: Comparison of four subset selection methods (component results are fused by linear combination and fusion results are evaluated by MAP)

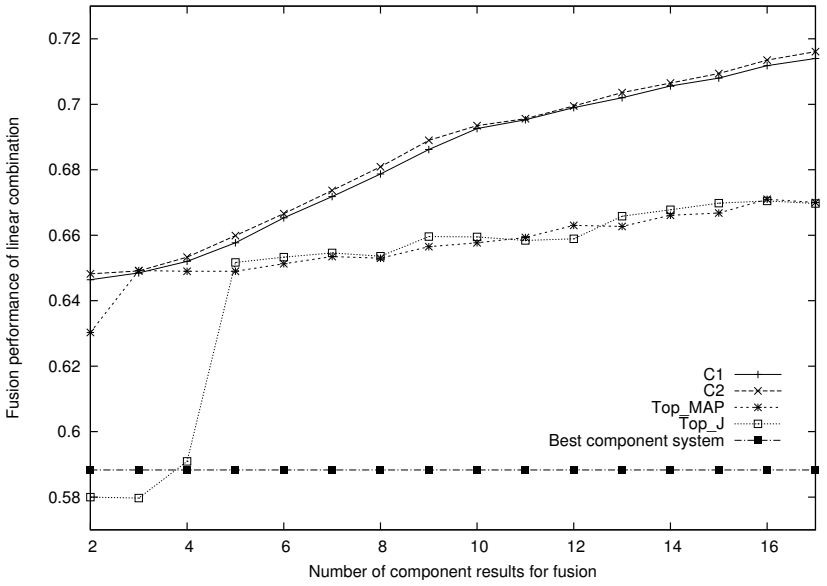

Figure 5: Comparison of four subset selection methods (component results are fused by linear combination and fusion results are evaluated by CAM)

50 times. The results presented in this section are the average of them.

In almost all the cases, CombMNZ is slightly worse than CombSum. Therefore, in the following we do not present CombMNZ's performance. Figs 2-5 present the performance of CombSum and linear combination.

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%

From these figures, we can see that C1 and C2 are better than Top_J and Top_MAP in most cases, whether CombSum or linear combination is used for fusion, and whether MAP or CAM is used for evaluation. Very often C1 and C2 are close. It shows that using a simple or a more sophisticated clustering method does not change fusion performance very much.

In all 51 runs submitted, the best performer is h2oloo.m5, with a MAP of 0.3832 and a CAM of 0.5883. Both C1 and C2 outperform it throughout from fusing 2 to 17 component systems. Obviously, h2oloo.m5 is always a participant in both C1 and C2. If we consider the situation of fusing 17 retrieval systems, then C1+CombSum achieves 0.4638 in MAP, and 0.7011 in CAM, which are better than h2oloo.m5 by 21.03% and 19.17%, respectively; C2+linear combination achieves 0.5036 in MAP, and 0.7161 in CAM, which are better than h2oloo.m5 by 31.42% and 21.72%, respectively. It is also noticeable that when fusing three to six systems, Top_MAP achieves the best performance with CombSum. On the other hand, the advantage of C1 and C2 is more prominent with linear combination throughout all different number of component retrieval systems.

The following Table 2 shows pairwise comparison results of all four methods on average of 16 groups of fusion (2-17 resulting lists). For example, the figure at column "C1:C2" and row "CombSum/MAP" means that 1.20% is the improvement rate of subset section methods C1 over C2 using CombSum for fusion and measured by MAP. A figure in bold indicates that the difference between the two methods is significant at the .05 level (paired samples t test). From Table 2 we can see that C1 and C2 are very close and better than the other two. C1 performs better than C2 when fused with CombSum, while C2 performs better than C1 when linear combination is used for fusion. Top_MAP is in the third place, while Top_J is the worst. The difference between either C1 or C2 and Top_J is always over 5%, while the difference in other situations is less than 5% apart from one case: C2 vs. Top_MAP fused by linear combination and measured by MAP.

5. Clustering & Subset Selection Analysis

In this section, we present some further observations and some analysis about clustering-based methods.

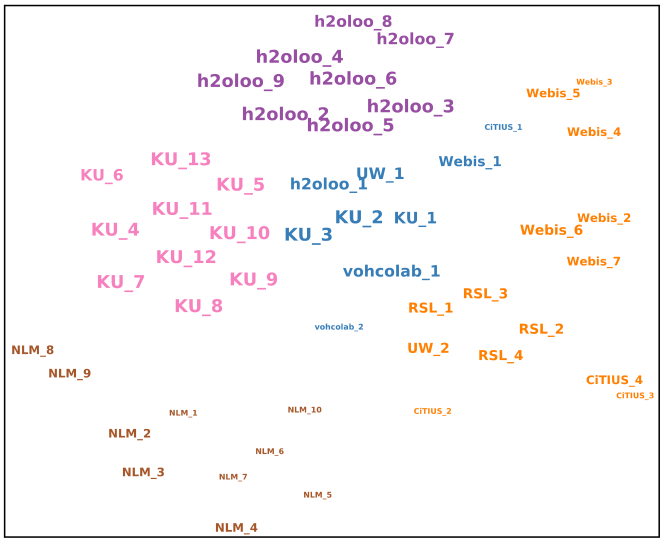

Figure 6: An clustering example of K-means with five clusters

5.1. Clustering Analysis

First let us look at clustering. Fig. 6 shows a clustering example of K-means with five clusters: all the resulting lists in a cluster are shown in the same colour, the distance between any two resulting lists represents their dissimilarity, and the size of the font represents the performance of the resulting list in MAP.

It seems that K-means does reasonably well in this example. However, we may observe that performance varies considerably across different clusters. As a matter of fact, the best in five clusters are h2oloo_5 (0.3832), KU_10 (0.3640), KU_3 (0.3122), RSL_4 (0.1913), and NLM_8 (0.1111), respectively. Three of them are much higher than the other two. Such an observation may be a positive evidence that generating more clusters is a good approach. If more clusters are generated, we can avoid picking some really bad ones. In this example, if we only choose three, then all selected runs are above 0.3 in MAP.

5.2. Subset Selection Analysis

In Section 4, we evaluated and compared four subset selection methods. Now for all the selected lists by each method, we calculate their average MAP and average pairwise distance between all the selected lists. See Table 3 for the detailed information. Distance values reflect the diversity of the chosen lists. From Table 3, we can see that Top_J and especially Top_MAP only choose those lists with top MAP values. When more lists are chosen, the average performance of resulting

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

lists in all four methods decrease. However, the decrease in C1 and C2 is much more quickly than it in Top_MAP and Top_J. On the other hand, higher distance values appear in all the cases for both C1 and C2 while that values are always lower for Top_MAP and Top_J. This give us a clear view of the four selection methods on two important aspects: performance and diversity. Top_MAP and Top_J only concern performance, they always choose top performers, but with less diversity, especially when a larger number of runs are selected. On the other hand, C1 & C2 have a balanced view about those two aspects. Compared with their counterparts Top_MAP and Top_J, more often they choose those runs with smaller MAP values but larger distance values on average. If we define a new measure $Combi$=0.5*MAP/Max_MAP+0.5*Dist/Max_Dist, where Max_MAP (0.377) and Max_Dist (4.629) are the maximal values observed, respectively, then we can find that in most cases C1 and C2 have large $Combi$ values than Top_MAP and Top_J do in almost all the cases except two. It can explain why C1 & C2 are more effective than Top_MAP and Top_J in most cases and on average.

6. 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. Experiments carried out with the Health Misinformation data set in TREC 2020 show that the proposed methods are very good. When fusing up to 17 retrieval systems, the proposed methods are better than the best component retrieval system by 20% to 30%, and they are also better than the state-of-the-art subset selection method by a clear margin. 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. Such results demonstrate that data fusion is a good approach for this Health Misinformation task.

In our future work, we plan to further 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. If some automatic performance estimation methods can be applied instead, then its usability can be improved.

References

[1] Marcoux Thomas, Agarwal Nitin, Narrative trends of covid-19 misinformation., in: Text2Story@ ECIR, 2021, pp. 77–80.

[2] Chandrasekaran Ranganathan, Mehta Vikalp, Valkunde Tejali, Moustakas Evangelos, Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study, Journal of medical Internet research 22 (2020) e22624.

[3] Abdelminaam Diaa Salama, Ismail Fatma Helmy, Taha Mohamed, Taha Ahmed, Houssein Essam H, Nabil Ayman, Coaid-deep: An optimized intelligent framework for automated detecting covid-19 misleading information on twitter, IEEE Access 9 (2021) 27840–27867.

[4] Ng Lynnette Hui Xian, Loke Jia Yuan, Analyzing public opinion and misinformation in a covid-19 telegram group chat, IEEE Internet Computing 25 (2020) 84–91.

[5] Leng Yan, Zhai Yujia, Sun Shaojing, Wu Yifei, Selzer Jordan, Strover Sharon, Zhang Hezhao, Chen Anfan, Ding Ying, Misinformation during the covid-19 outbreak in china: Cultural, social and political entanglements, IEEE Transactions on Big Data 7 (2021) 69–80.

[6] Zhou Cheng, Xiu Haoxin, Wang Yuqiu, Yu Xinyao, Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on covid-19, Information Processing & Management 58 (2021) 102554.

[7] Shahi Gautam Kishore, Dirkson Anne, Majchrzak Tim A, An exploratory study of covid-19 misinformation on twitter, Online social networks and media 22 (2021) 100104.

[8] Choudrie Jyoti, Banerjee Snehasish, Kotecha Ketan, Walambe Rahee, Karende Hema, Ameta Juhi, Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study, Computers in Human Behavior 119 (2021) 106716.

[9] Roberts Kirk, Alam Tasmeer, Bedrick Steven, Demner-Fushman Dina, Lo Kyle, Soboroff Ian, Voorhees Ellen, Wang Lucy Lu, Hersh William R, Trec-covid: rationale and structure of an information retrieval shared task for covid-19, Journal of the American Medical Informatics Association 27 (2020) 1431–1436.

[10] Charles L.A.Clarke, Maria Maistro, Saira Rizvi, Mark D.Smuckerand Guido Zuccon, Overview of the TREC 2020 health misinformation track, in: Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, 2020.

[11] Budíková Petra, Batko Michal, Zezula Pavel, Fusion strategies for large-scale multi-modal image retrieval, in: Transactions on Large-Scale Data-and Knowledge-Centered Systems

XXXIII, Springer, 2017, pp. 146–184.
[12] Kato Sosuke, Shimizu Toru, Fujita Sumio, Sakai Tetsuya, Unsupervised answer retrieval with data fusion for community question answering, in: Asia Information Retrieval Symposium, Springer, 2019, pp. 10–21.
[13] Roostaee Meysam, Sadreddini Mohammad Hadi, Fakhrahmad Seyed Mostafa, An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes, Information Processing & Management 57 (2020) 102150.
[14] Smeaton Alan F, O'Connor Edel, Regan Fiona, Multimedia information retrieval and environmental monitoring: Shared perspectives on data fusion, Ecological informatics 23 (2014) 118–125.
[15] Juárez-González Antonio, Montes-y-Gómez Manuel, Villaseñor-Pineda Luis, Pinto-Avendaño David, Pérez-Coutiño Manuel, Selecting the n-top retrieval result lists for an effective data fusion, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2010, pp. 580–589.
[16] Zhao Yuehua, Da Jingwei, Yan Jiaqi, Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches, Information Processing & Management 58 (2021) 102390.
[17] Wang Zuhui, Yin Zhaozheng, Argyris Young Anna, Detecting medical misinformation on social media using multimodal deep learning, IEEE Journal of Biomedical and Health Informatics 25 (2020) 2193–2203.
[18] Zhang Qiang, Cook Jonathan, Yilmaz Emine, Detecting and forecasting misinformation via temporal and geometric propagation patterns., in: ECIR (2), 2021, pp. 455–462.
[19] Lee Nayeon, Li Belinda Z, Wang Sinong, Fung Pascale, Ma Hao, Yih Wen-tau, Khabsa Madian, On unifying misinformation detection, arXiv preprint arXiv:2104.05243 (2021).
[20] Ginsca Alexandru L, Popescu Adrian, Lupu Mihai, Credibility in information retrieval, Foundations and Trends in Information Retrieval 9 (2015) 355–475.
[21] B. J. Fogg, H. Tseng, The elements of computer credibility, in: M. G. Williams, M. W. Altom (Eds.), Proceeding of the CHI '99 Conference on Human Factors in Computing Systems: The CHI is the Limit, Pittsburgh, PA, USA, May 15-20, 1999, ACM, 1999, pp. 80–87.
[22] S. Tseng, B. J. Fogg, Credibility and computing technology, Commun. ACM 42 (1999) 39–44.
[23] M. J. Metzger, A. J. Flanagin, Information in online environments: the use of cognitive heuristics, Journal of Pragmatics 59 (2013) 210–220.
[24] M. Viviani, G. Pasi, Credibility in social media: opinions, news, and health information - a survey, WIREs Data Mining Knowl. Discov. 7 (2017).
[25] Song Shijie, Zhang Yan, Yu Bei, Interventions to support consumer evaluation of online health information credibility: A scoping review, International Journal of Medical Informatics 145 (2021) 104321.
[26] Wu Shu, Liu Qiang, Liu Yong, Wang Liang, Tan Tieniu, Information credibility evaluation on social media, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.
[27] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, A. Alamri, A credibility analysis system for assessing information on twitter, IEEE Transactions on Dependable and Secure Computing

15 (2016) 661–674.
[28] Wu Lianwei, Rao Yuan, Nazir Ambreen, Jin Haolin, Discovering differential features: Adversarial learning for information credibility evaluation, Information Sciences 516 (2020) 453–473.
[29] Fox Edward A, Koushik M Prabhakar, Shaw Joseph, Modlin Russell, Rao Durgesh, et al., Combining evidence from multiple searches, in: The first text retrieval conference (TREC-1), 1993, pp. 319–328.
[30] Cormack Gordon V, Clarke Charles LA, Buettcher Stefan, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 758–759.
[31] Wu Shengli, Linear combination of component results in information retrieval, Data & Knowledge Engineering 71 (2012) 114–126.
[32] Clipa Teofan, Di Nunzio Giorgio Maria, A study on ranking fusion approaches for the retrieval of medical publications, Information 11 (2020) 103.
[33] de Herrera Alba G Seco, Schaer Roger, Markonis Dimitrios, Müller Henning, Comparing fusion techniques for the imageclef 2013 medical case retrieval task, Computerized Medical Imaging and Graphics 39 (2015) 46–54.
[34] Mourão André, Martins Flávio, Magalhaes Joao, Multimodal medical information retrieval with unsupervised rank fusion, Computerized Medical Imaging and Graphics 39 (2015) 35–45.
[35] Fernández-Pichel Marcos, Losada David E, Pichel Juan C, Elsweiler David, Citius at the trec 2020 health misinformation track 1266 (2020).
[36] Pradeep Ronak, Ma Xueguang, Zhang Xinyu, Cui Hang, Xu Ruizhou, Nogueira Rodrigo, Lin Jimmy, H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine, Corpus 5 (2020) d2.
[37] Leonard David, Lillis David, Zhang Lusheng, Toolan Fergus, Collier Rem W, Dunnion John, Applying machine learning diversity metrics to data fusion in information retrieval, in: European Conference on Information Retrieval, Springer, 2011, pp. 695–698.
[38] Wu Shengli, McClean Sally, Performance prediction of data fusion for information retrieval, Information processing & management 42 (2006) 899–915.
[39] Bradley Paul S, Fayyad Usama M, Refining initial points for k-means clustering., in: ICML, volume 98, Citeseer, 1998, pp. 91–99.
[40] Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D. Smucker, Guido Zuccon, Overview of the trec 2019 decision track, in: TREC, volume 1250, Special Publication, 2019.
[41] Fox Edward A, Shaw Joseph A, Combination of multiple searches, NIST special publication SP 243 (1994).