Fusing blog opinion retrieval results for better effectiveness

Shengli Wu

School of Computing and Mathematics, University of Ulster, UK

Data fusion in information retrieval

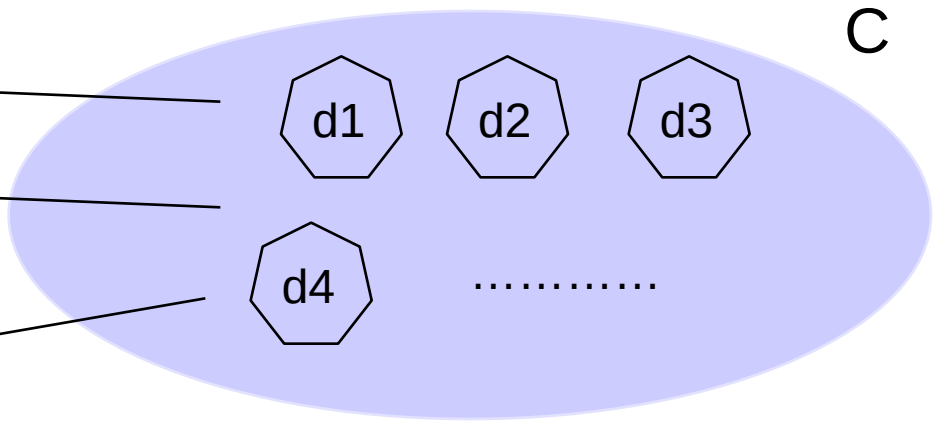
Suppose for the same document collection C and a given query Q, we have a group of component results $R_i(1 \le i \le n)$, each of which is from a different retrieval system:

$R_{1:}\ d_{11}, d_{12}, \ldots, d_{1m}$

$R_{2:}\ d_{21}, d_{12}, \ldots, d_{2m}$

.....

$R_{n:}\ d_{n1}, d_{n2}, \ldots, d_{nm}$



fusing them to get $R_{f:}\ d'_1, d'_2, \ldots, d'_m$

We hope that the fused result is more effective than component results.

Several Data Fusion Methods

Component results R_i is composed of a ranked list of documents, each document has a relevance score:

$R_{1:}$ $d_1(0.8)$, d_3 (0.5), $d_4(0.2)$

$R_{2:}$ $d_2(0.6)$, $d_4(0.5)$, $d_3(0.4)$

CombSum: $R_{f\text{-}sum:}$ $d_3(0.5+0.4)$, d_1 $(0.8+0)$, $d_4(0.2+0.5)$, $d_2(0+0.6)$

CombMNZ: $R_{f\text{-}mnz:}$ $d_3(0.9*2)$, $d_4(0.7*2)$, $d_1(0.8*1)$, d_2 $(0.6*1)$

Linear Combination: e.g., assign a weight 2 to R_1 and a weight 3 to R_2

$R_{f\text{-}lc:}$ $d_3(0.5*2+0.4*3=2.2)$, $d_4(0.2*2+0.5*3=1.9)$, $d_2(0.6*3=1.8)$, d_1 $(0.8*2=1.6)$

The linear combination method

- The linear combination method is a very flexible method since different weights can be assigned to different systems;
- It is especially useful when either or both effectiveness of component results and similarity between different component results vary.

Weights assignment

- Obviously, how to assignment weights is an important issue;
- Some training data are required for doing it.
- Method 1: performance level weighting (LCP, only effectiveness is considered).
- Method 2: performance square weighting (LCP2, only effectiveness is considered)
- Method 3: trained by linear regression (LCR, both effectiveness and similarity between results are considered)

Score normalization

- Borda count: for a ranked list of 1000 document, the first is assigned a score of 1000, the second 999, and so on, …, the last is assigned a sore of 1.
- The fitting method, an improved version of the zero-one linear score normalization: set appropriate min and max, let (0<min<max<1), n_score=(raw_score-min)/(max-min)

Information of the data set used

- The “Blog06” test collection
- Used in the TREC 2008 blog track (opinion retrieval)
- Total uncompressed size 148GB
- Number of unique blogs 100,649
- Number of feeds fetches 753,681
- Number of permalinks 3,215,171
- Number of homepages 324,880
- 5 standard baselines and 191 runs submitted by 19 groups

Measures used

- Average precision over all relevant documents
- Recall-level precision
- Precision at 10 document levels
- Reciprocal rank

Experimental setting

- Divide all 150 queries into three groups. The first group includes query 1, 4, 7, 10,…; the second group includes query 2, 5, 8, 11,…; the third group includes query 3, 6, 9, 12, ….
- Using one group as training data, and the other two as test data. Repeat for three times.
- From all 191 runs submitted, randomly choose 5, 10, 15,…, 60 runs for the experiment. For each given number, repeat for 200 times.
- The experimental result is the average of them.

Experimental result (MAP, Borda normlization)

Num.	Best	C'Sum	C'MNZ	LCP	LCP2	LCR
5	0.378	0.416+	0.410+	0.425+	0.428+	**0.441+**
10	0.403	0.447+	0.438+	0.454+	**0.458+**	**0.454+**
15	0.417	0.458+	0.447+	0.464+	**0.467+**	**0.464+**
20	0.431	0.467+	0.456+	0.472+	0.476+	**0.478+**
25	0.442	0.472+	0.460+	0.477+	0.481+	**0.484+**
30	0.449	0.473+	0.461+	0.478+	0.483+	**0.490+**
35	0.451	0.476+	0.464+	0.480+	0.483+	**0.489+**
40	0.460	0.475+	0.463	0.480+	0.484+	**0.490+**
45	0.468	0.478+	0.466	0.483+	0.487+	**0.495+**
50	0.471	0.480+	0.468	0.485+	0.489+	**0.497+**
55	0.473	0.480	0.468	0.484+	0.489+	**0.494+**
60	0.481	0.481	0.469−	0.486+	0.491+	**0.501+**
Ave.	0.444	0.467	0.456	0.472	0.476	**0.481**
		5.28%	2.77%	6.70%	7.42%	8.52%

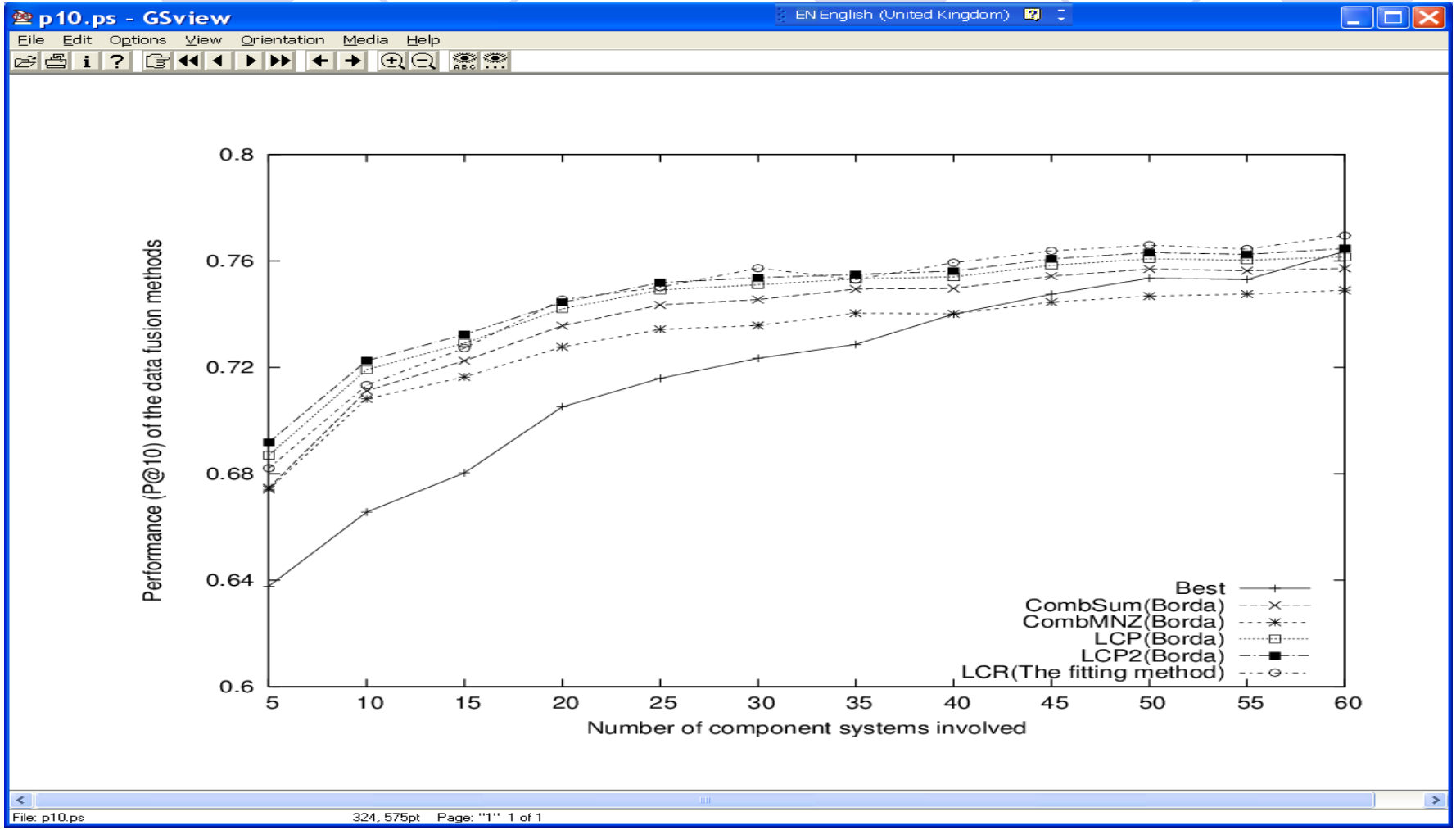
Experimental result (MAP, fitting normlization)

Num.	Best	C'Sum	C'MNZ	LCP	LCP2	LCR
5	0.378	0.417+	0.412+	0.423+	0.425+	**0.425+**
10	0.403	0.447+	0.440+	0.451+	0.453+	**0.459+**
15	0.417	0.455+	0.448+	0.458+	0.461+	**0.473+**
20	0.431	0.462+	0.456+	0.465+	0.468+	**0.488+**
25	0.442	0.467+	0.460+	0.470+	0.473+	**0.493+**
30	0.449	0.467+	0.460+	0.470+	0.473+	**0.498+**
35	0.451	0.468+	0.462+	0.470+	0.473+	**0.497+**
40	0.460	0.468+	0.462	0.471+	0.474+	**0.501+**
45	0.468	0.471+	0.464−	0.473+	0.477+	**0.506+**
50	0.471	0.472	0.466−	0.474	0.477+	**0.507+**
55	0.473	0.472	0.466−	0.474	0.478	**0.508+**
60	0.481	0.473−	0.466−	0.475−	0.479	**0.511+**
Ave.	0.444	0.462	0.455	0.464	0.468	**0.489**
		4.08%	2.62%	4.69%	5.41%	10.26%

Experimental result (MRR, fitting normlization)

Num.	Best	C'Sum	C'MNZ	LCP	LCP2	LCR
5	0.804	0.838+	0.837+	**0.840+**	**0.839+**	**0.830+**
10	0.824	0.857+	0.857+	0.858+	**0.858+**	**0.858+**
15	0.835	0.864+	0.865+	0.864+	0.865+	**0.870+**
20	0.851	0.871+	0.873+	0.871+	0.873+	**0.882+**
25	0.859	0.874+	0.877+	0.874+	0.875+	**0.885+**
30	0.863	0.873+	0.877+	0.873+	0.875+	**0.891+**
35	0.865	0.873+	0.878+	0.874+	0.876+	**0.886+**
40	0.871	0.874	0.879+	0.874	0.877+	**0.890+**
45	0.877	0.876	0.881+	0.876	0.878	**0.892+**
50	0.880	0.879	0.884+	0.879	0.881	**0.891+**
55	0.881	0.879	0.884	0.878	0.880	**0.890+**
60	0.888	0.880	0.885	0.880−	0.881−	**0.896+**
Ave.	0.858	0.870	0.873	0.870	0.871	**0.880**
		1.38%	2.35%	1.40%	1.54%	2.56%

Experimental result (p@10, fitting normlization)



Observations

In summary, one major observation from this study is: in most cases, the combination of the fitting method for score normalization and multiple linear regression for weights assignment is the most effective approach, especially when a relatively large number of component systems are fused.

Other Observations (1)

1) CombSum and CombMNZ are always close. Most of the time CombSum is a little better than CombMNZ. Sometimes the difference is significant, sometimes it is not.

2) With very few exceptions, LCP2 is always a little better than LCP. The difference between them is very often significant.

3) On average, LCP2 is the second best method in the experiment. It consistently outperforms the best component system when 35 or less systems are fused.

Other Observations (2)

4) When a relatively small number (say, 5 or 10) of component systems are fused, then all data fusion methods outperform the best component system by a clear margin.
5) Compared with the best system, all the data fusion methods are more effective on average. However, the improvement rate varies when different metrics are used. The most favourable metric is AP, followed by RP and P@10, while RR is the least favourable.
6) For CombSum, CombMNZ, LCP, and LCP2, Borda is better; for LCR, the fitting method is better.

Note

A longer version of this paper has been accepted for publication.

Expert Systems with Applications :
Applying the data fusion technique to blog opinion retrieval
Shengli Wu

http://www.sciencedirect.com/science/article/pii/S0957417411011262