Condorcet fusion for blog opinion retrieval

Shengli Wu

School of Computer Science and telecommunication engineering
Jiangsu University, China

Data fusion in information retrieval

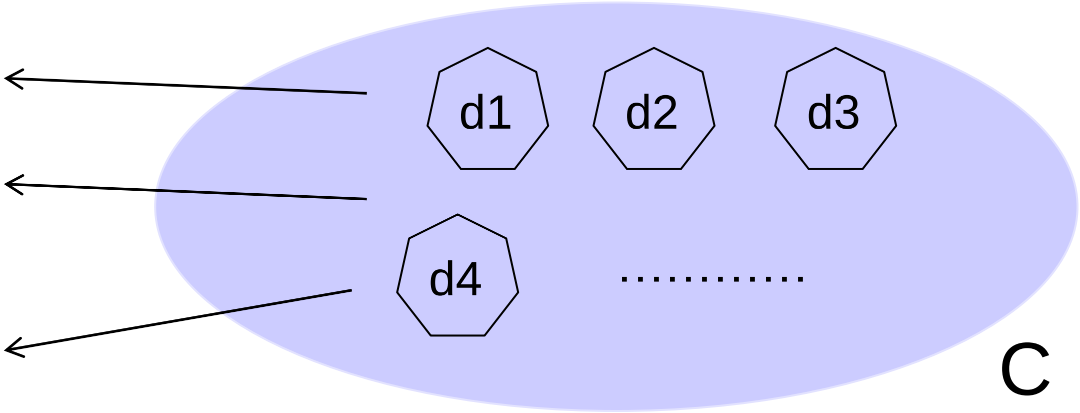
Suppose for the same document collection C and a given query Q, we have a group of component results $R_i(1 \le i \le n)$, each of which is from a different retrieval system:

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

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

……

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



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)

Condorcet fusion

- Compare every pair of documents : d_1 and d_2. If d_1 is favoured than d_2 by more information retrieval systems, then we rank d_1 ahead of d_2; If d_2 is favoured than d_1 by more information retrieval systems, then we rank d_2 ahead of d_1 .
- It is possible to obtain a draw between two or more documents. How to handle this is important. Different variants of the Condorcet method deal with this in different ways. A easy solution is taken in this study.
- A final ranking of all the documents can be generated by considering all different pairs.

Property of Condorcet

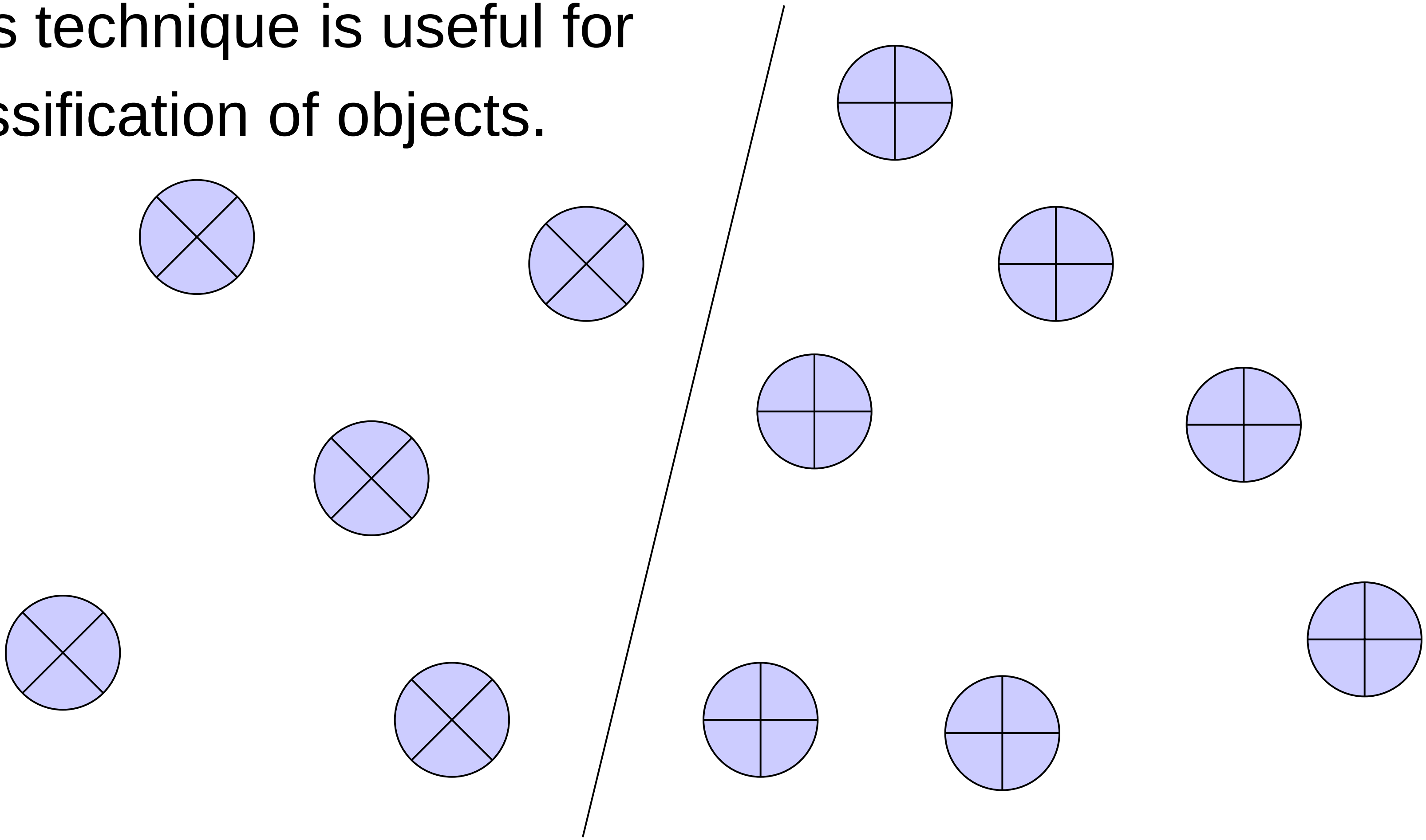
- Ideally, when all the information retrieval systems are equally effective and equally similar to each other, than Condorcet is a very effective data fusion method.
- When the condition move away from ideal, then Condorcet worse off very quickly than other methods such as CombSum and CombMNZ.
- In such a situation, to use weighted Condorcet becomes desirable.

Weights assignment

- Condorcet is quite different from other score-based data fusion methods.
- The weights assignment method used for the linear combination method cannot be used directly for weighted Condorcet.
- We find that linear discriminate analysis is an effective technique for this task.

Linear Discriminate analysis

This technique is useful for classification of objects.



Methodology (considering the table)

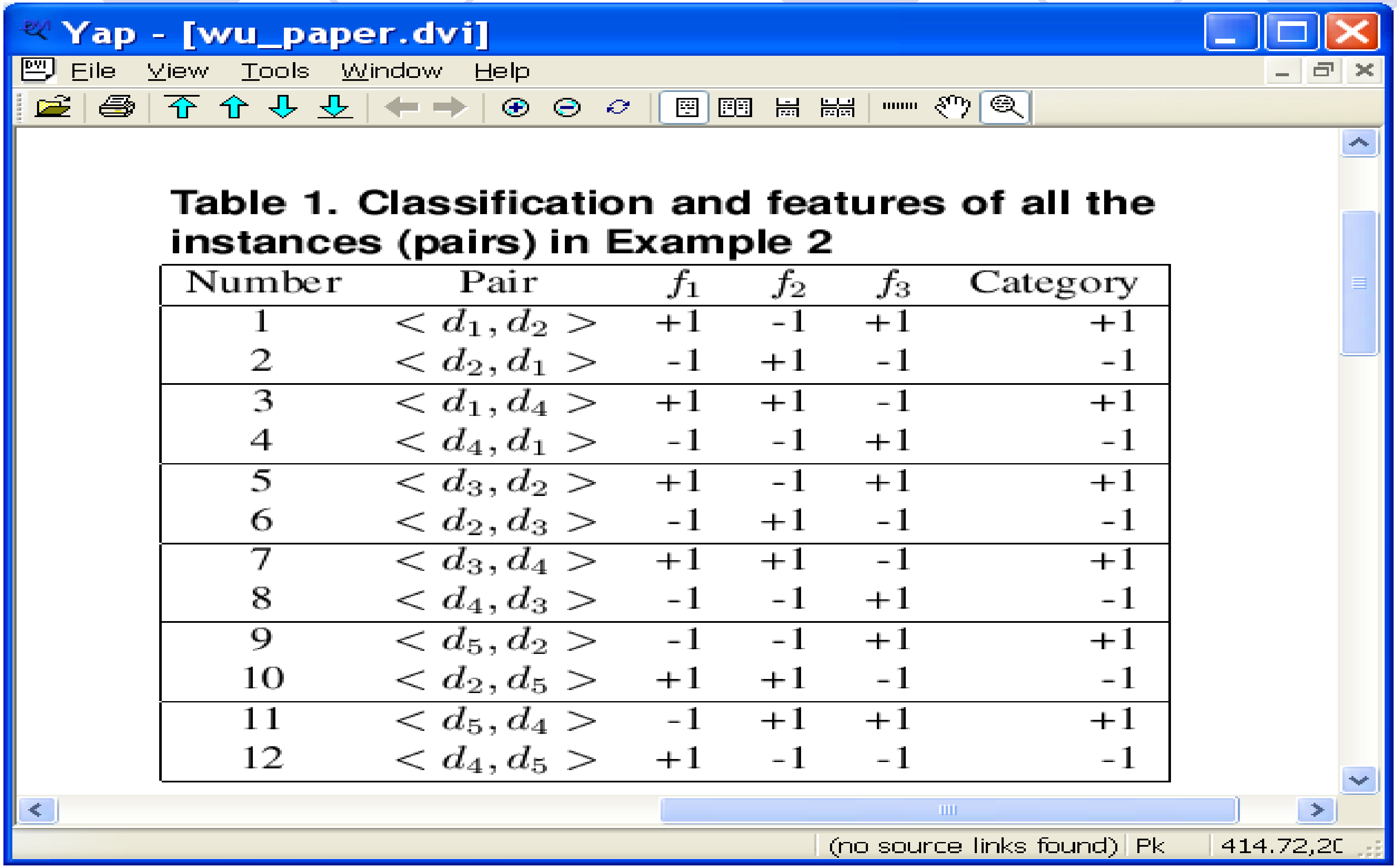


Table 1. Classification and features of all the instances (pairs) in Example 2

Number	Pair	f_1	f_2	f_3	Category
1	$< d_1, d_2 >$	+1	-1	+1	+1
2	$< d_2, d_1 >$	-1	+1	-1	-1
3	$< d_1, d_4 >$	+1	+1	-1	+1
4	$< d_4, d_1 >$	-1	-1	+1	-1
5	$< d_3, d_2 >$	+1	-1	+1	+1
6	$< d_2, d_3 >$	-1	+1	-1	-1
7	$< d_3, d_4 >$	+1	+1	-1	+1
8	$< d_4, d_3 >$	-1	-1	+1	-1
9	$< d_5, d_2 >$	-1	-1	+1	+1
10	$< d_2, d_5 >$	+1	+1	-1	-1
11	$< d_5, d_4 >$	-1	+1	+1	+1
12	$< d_4, d_5 >$	+1	-1	-1	-1

Information of the data sets 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 3, 4, 5,…, 10 runs for the experiment. For each given number, repeat for 200 times.

The experimental result is the average of them.

Performance of data fusion methods (average precision, Figure 1)

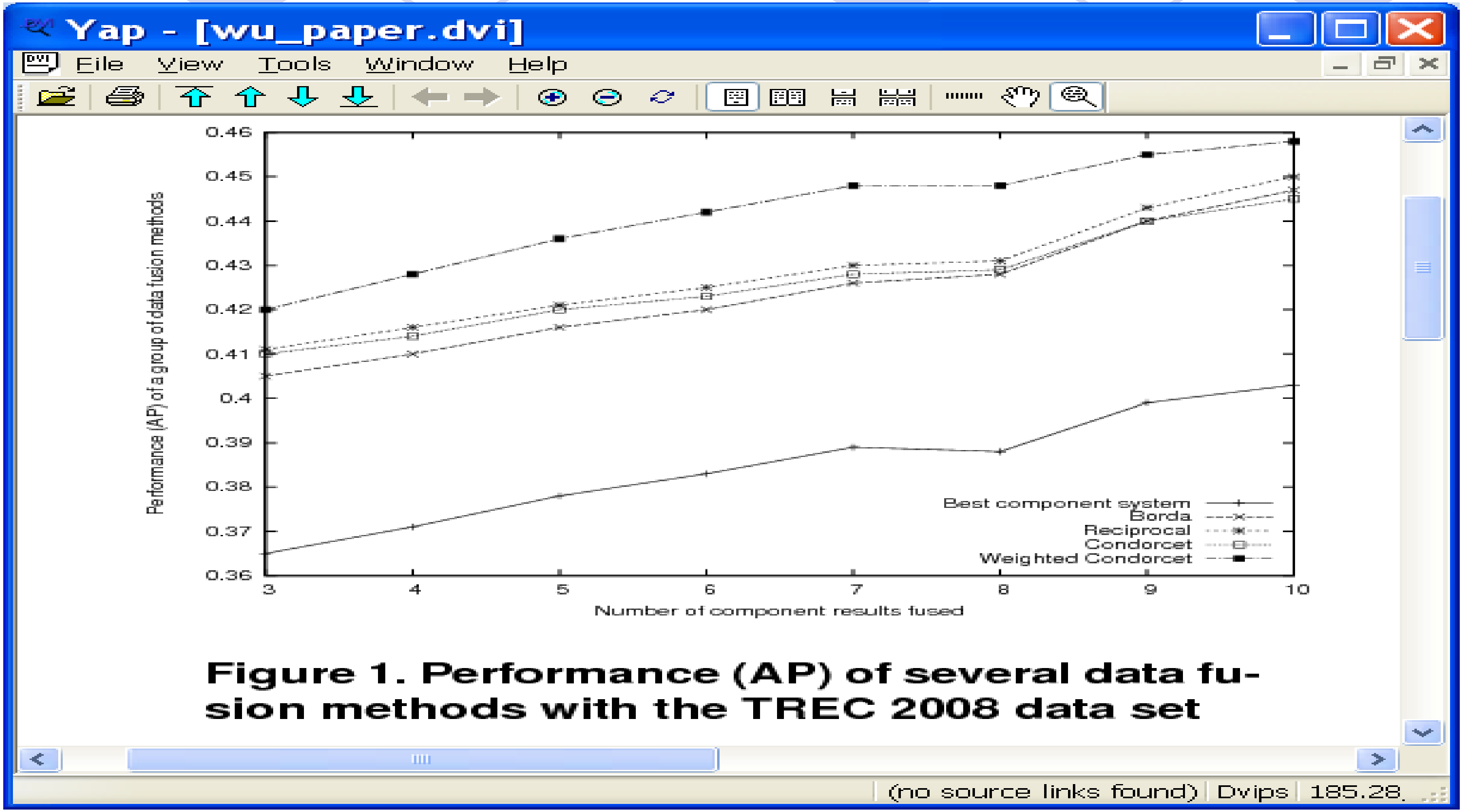


Figure 1. Performance (AP) of several data fusion methods with the TREC 2008 data set

Performance of data fusion methods (recall-level precision, Figure 2)

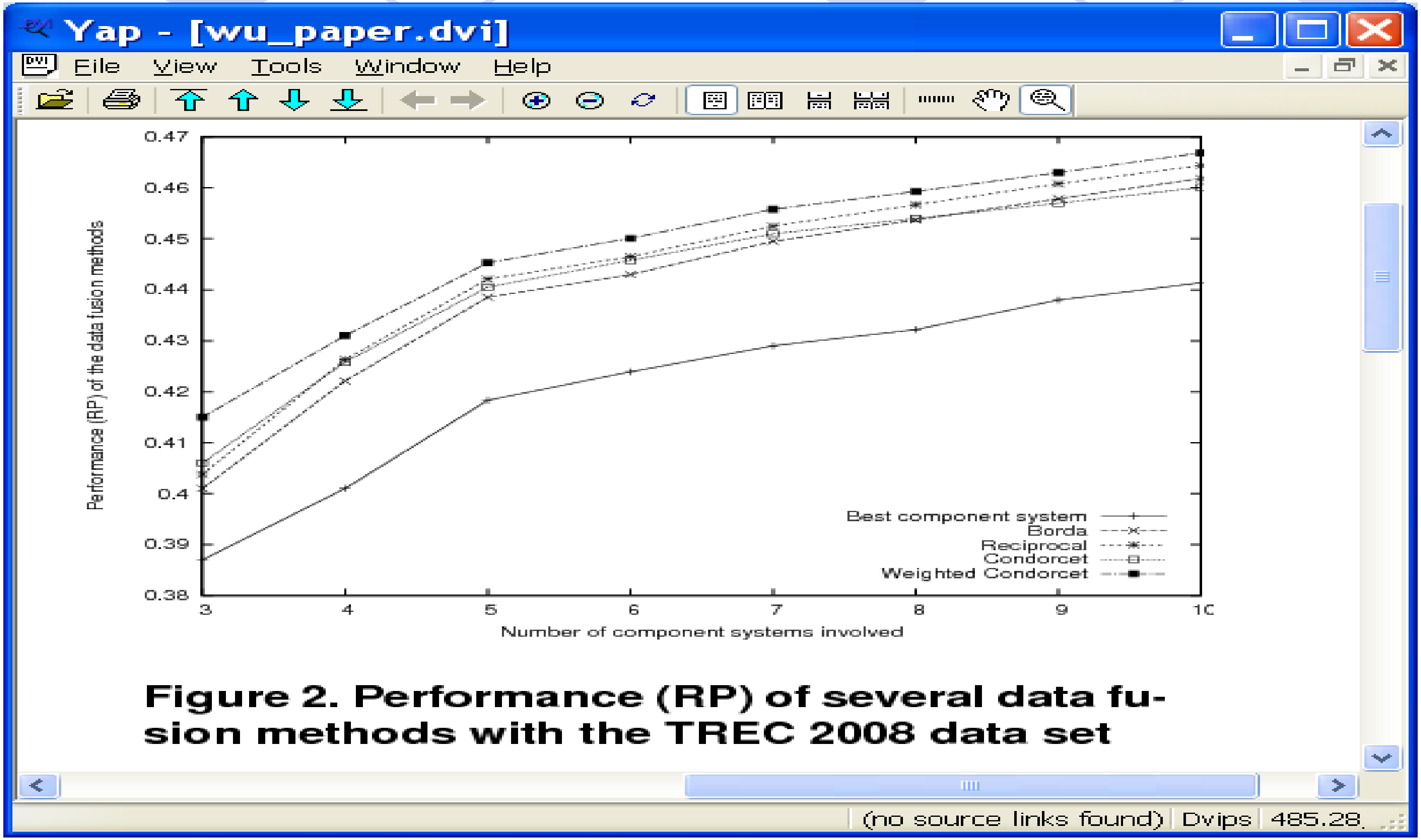


Figure 2. Performance (RP) of several data fusion methods with the TREC 2008 data set

Performance of data fusion methods (precision at 10 document level, Figure 3)

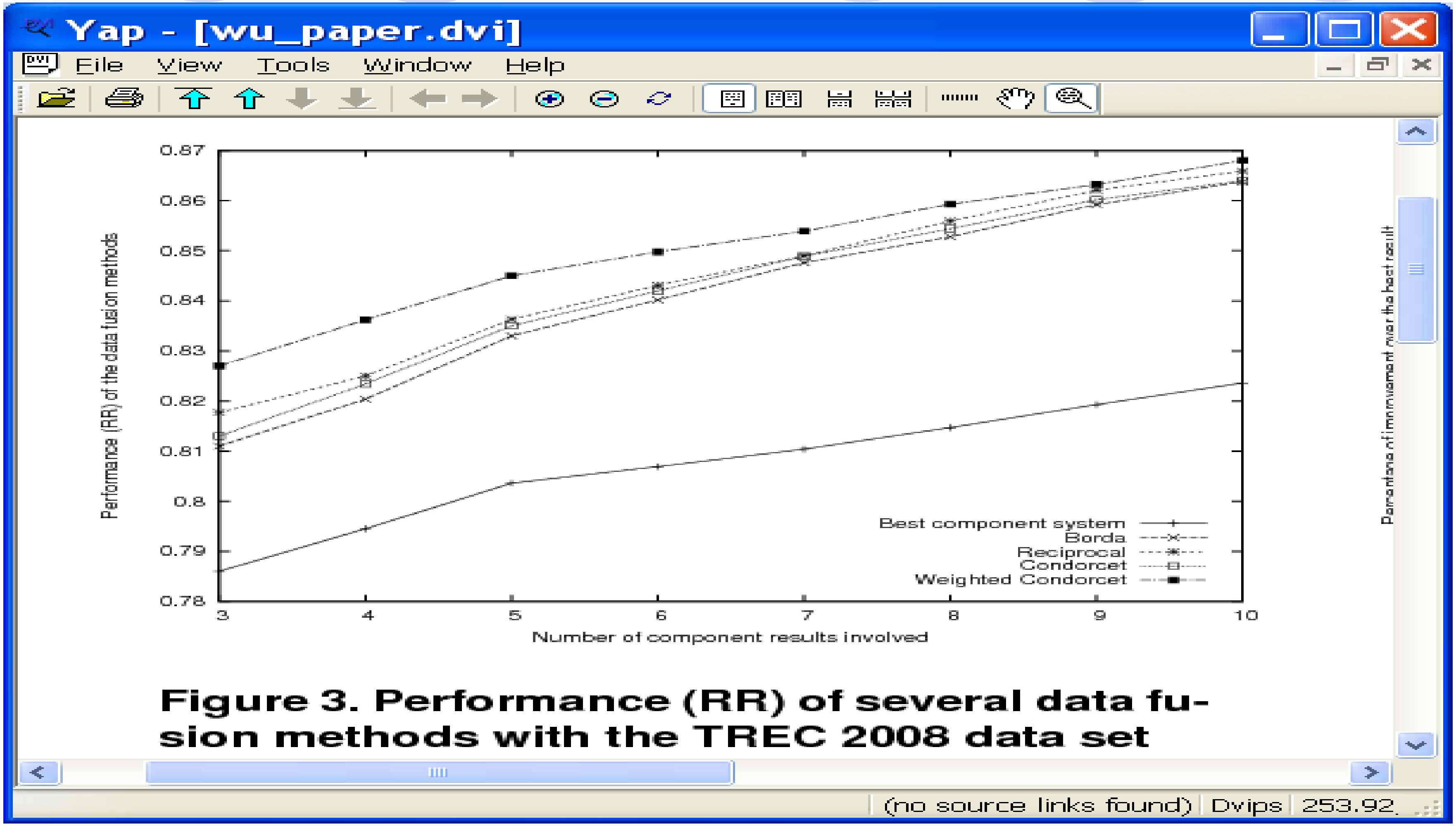


Figure 3. Performance (RR) of several data fusion methods with the TREC 2008 data set

Improvement of data fusion methods (over best component result, Figure 4)

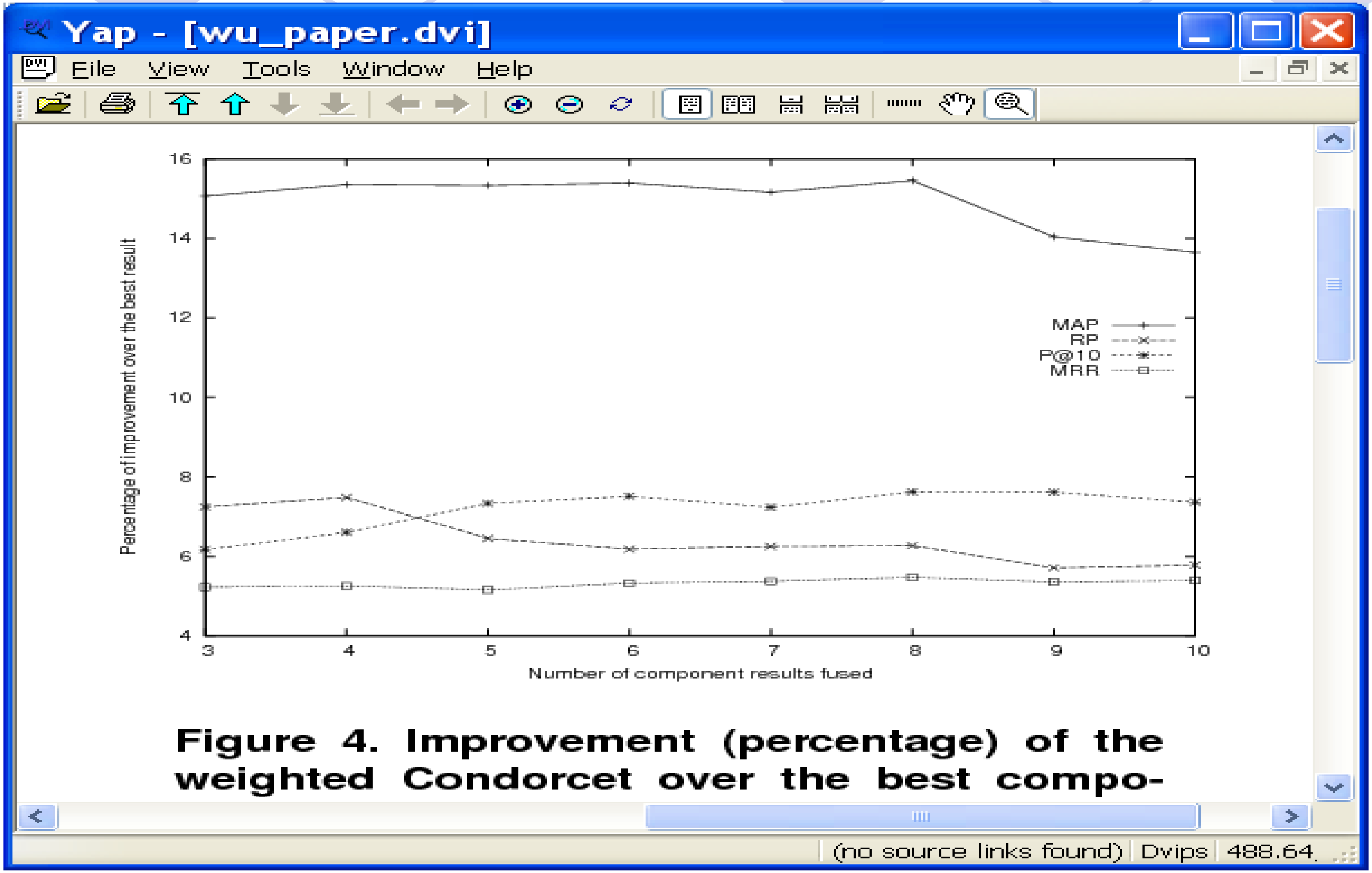

Yap - [wu_paper.dvi]
File
View
Tools
Window
Help
16
14
12
10
8
6
4
3
4
5
6
7
8
9
10
Percentage of improvement over the best result
Number of component results fused
MAP
RP
P@10
MRR

Figure 4. Improvement (percentage) of the weighted Condorcet over the best compo-

(no source links found) Dvips 488.64

Conclusions

- Data fusion can be helpful for improving effectiveness if used properly. All the data fusion methods involved perform better than the best component result.
- LDA is a good method of weights assignment for the weighted Condorcet. Weighted Condorcet is more effective than other data fusion methods. On the other hand, it requires training for weights assignment, while other methods do not need this.

More in a journal paper

- Shengli Wu, **The weighted Condorcet fusion in information retrieval,** Information Processing & Management, available online

- http://dx.doi.org/10.1016/j.ipm.2012.02.007