Does Overlap Mean Relevance?

João Ferreira
Instituto Superior de Eng. de Lisboa
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
Jose.Delgado@tagus.ist.utl.pt

Abstract: The work developed in this paper focuses on the overlap of relevant retrieved documents obtained from different combination of systems and components. Study of the impact on information retrieval systems performance is also carried out.

Keywords: Information Retrieval, combination, overlap, relevance

1. Introduction

We argue in this paper that combine and integrate combination components are the key question in combination Information Retrieval (IR). Two of the most common ways combination can be applied is at *retrieval time* (i.e. combination components are integrated to produce one set of results) or *after retrieval* (i.e. multiple sets of results, produced by combination components applied in parallel, are merged after retrieval time). In post-retrieval combination, which is the combination approach taken in this paper, two of the most common combination formulas are *Similarity Merge* [1,2] and *Weighted Sum* [3,4,5,6].

Furthermore, some researchers observed that the inclusion of a "weak" component into the combination pot still results in strong performance gain, which suggests the possibility that combination can produce the whole greater than the sum of is parts. The potential of combination to leverage the strengths of its components while minimizing their weaknesses is not only promising in its own right, but offers a novel perspective of IR that relaxes the research goal of discovering just one best retrieval strategy. One explanation for this is the overlap of results of different systems. In this paper we explore this subject and show that *overlap is beneficial to improve the relevance of information retrieval systems*.

We divide the paper in 4 sections. In Section 1 we introduce the context and thesis for our research. Section 2 describes our experience. Section 3 discusses the overlap results and Section 4 presents the main conclusions.

2. Experience

The challenges and opportunities of Web IR motivated us to consider the possibilities of leveraging and combining Multiple Source of Evidence. In this paper, we combine methods that leverage text, hyperlink, and Web directory information to see how overlap of combination results can improve retrieval performance.

We combine information from different data sources, namely from: (1)- 36 VSM (Vector Space Model) systems; (2)- 6 HITS (Link-Based Retrieval) systems; (3)- 24 TM (Term Match) systems.

See [7] for more details. We use OpenFts <http://openfts.sourceforge.net/> system with the combination of **Similarity Merge (SM) (1),**

$$CS = (\sum NS_i) * \frac{olp}{m(i)} \quad (1) \qquad NS_i = (S_i - S_{min}) / (S_{max} - S_{min}) \quad (2) \qquad CS = \sum(w_i * RS_i), \quad (3)$$

CS=combination score of a document; NSi = normalized score of a document by system I; Olp = number of systems that retrieved a given document; m(i) = number of systems in a method to which system i belongs. The normalized document score, NS_i, (2) is computed by Lee's min-max formula [2,8], where S_i is the retrieval score of a given document and S_{max} and S_{min} are the maximum and minimum document scores by system *i*. The ***Weighted Rank Sum*** **(WRS)** formula (3), which uses rank-based

In *Proceedings of WWW/Internet 2004 (IADIS) Conference (WI2004)*.

scores (e.g. 1/rank) in place of document scores of WS formula, w_i = weight of system i; RS_i = rank-based score of a document by system i;

As data we use the WT10g collection [10], which is a ten-gigabyte subset of the 1997 Web crawl by the Internet Archive, consists of 1.7 million Web documents, 100 TREC queries (topics 451-550), and official NIST relevance judgments.

Because classification systems for the Web lack an ideal Web directory, we use Yahoo <http://yahoo.com> due to its popularity and size.

Systems notation is: v$query$index$phare$feedback; h$address$v*c10; t$#cat$wt10g-index$phrase

Where; query or address is: p-short;m-medium;l-long; phrase or feedback is: 1-yes; 0-no; wt10g-index is: 0-body text no phrase; 1-body text w/ phrase; 2 document no phrase; 3 document w/ phrase; index is: c-body text; t- header, d- document.

3. Overlap Results

Internal systems combination results [7] suggest that HITS systems have the most to gain by combinations due to their diverse solution spaces. One way to confirm such hypothesis is to examine the degree of overlap in relevant documents retrieved by HITS systems.

Tópicos 451-500								*Tópicos 501-550*							
Sistema	RRN	VSM	HITS	TM	V-H	V-T	H-T	Sistema	RRN	VSM	HITS	TM	V-H	V-T	H-T
vmc10	1340	0	-	-	0	0	-	vlc10	1963	1			1	1	
vmc11	1330	0	-	-	0	0	-	vlc00	1931	3			2	2	
vmc00	1324	0	-	-	0	0	-	vlc11	1917	3			3	3	
t110	948	-	-	0	-	0	0	t220	1295			0		0	0
t120	948	-	-	0	-	0	0	t210	1292			0		0	0
t111	943	-	-	0	-	0	0	t211	1288			3		1	2
hpl	724	-	90	-	6	-	52	hpl	1162		157		3		42
hpm	732	-	35	-	0	-	9	hpm	1043		37		3		17
hpp	633	-	50	-	4	-	6	hpp	965		69		3		19

Table 1: Lists the total number of relevant documents (RRN) as well as the number of relevant documents uniquely retrieved by a system within a given method (e.g. VSM, HITS, TM).

Table 1 describes the degree of overlap in relevant documents retrieved. The VSM, HITS, and TM columns indicate that the solution spaces for HITS systems overlap much less than those of VSM or TM systems. More specifically, the unique contributions of the top 3 HITS systems, which are considerably larger than those of the top 3 VSM or TM systems, imply that the HITS method has the most to gain by combination.

Examination of the external systems combination results shows that they also deserve overlap analysis. Larger numbers in the H-T column of Table 1 indicate the large potential gain for HITS-TM combination. Different kinds of overlap analysis are required to explain the fact that VSM-HITS combination results were closer to the upper bound defined by the best VSM system while VSM-TM combination results fell more towards the middle of the upper and lower bound defined by the best VSM and TM systems.

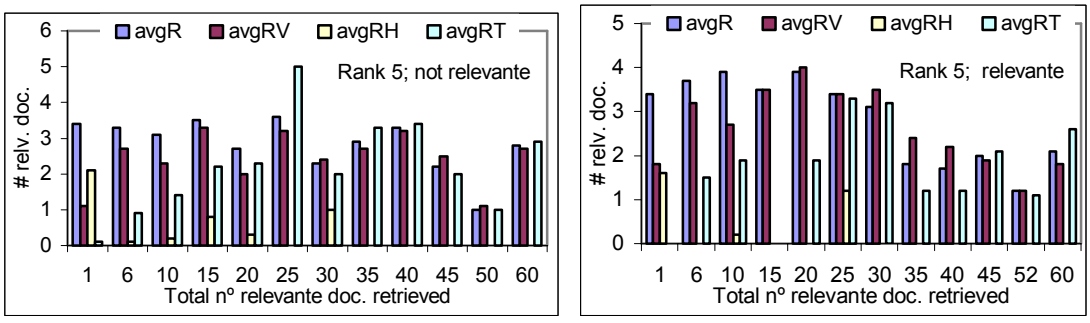

Figure 1: Overlap statistics for all systems at rank 5 for topics 451-500.

avgR = average rank in a partition; avgRV = average rank of VSM system results; avgRH = average rank of HITS system results; avgRT = average rank of TM system results in a given partition;

Figure 1 depicts overlap statistics, which list the frequency and relevance percentage of the overlapped documents (i.e. documents retrieved by multiple systems). These show a fairly even number of overlap counts for VSM and TM but hardly any overlap count for HITS at high ranks [5]. Even at lower ranks, documents retrieved by many HITS systems are small in numbers compared to VSM and TM. Documents retrieved by VSM systems are much more likely to influence the VSM-HITS combined

results than documents retrieved by HITS systems, due to fact that combination formulas favor overlapped documents. Consequently, the VSM-HITS combination results are closer to the VSM (rather than the HITS) baseline. When VSM and TM system results are combined, however, documents retrieved by either system get the overlap boost and the results of VSM systems get degraded by the large number of non-relevant documents with high overlap in TM systems, as shown in figure 1 (see [10] for more details).

Figure 2 shows that combining all VSM systems (Fv) for topics 451-500 could increase the optimum average precision of the best VSM systems from 0.6398 to 0.7555 by introducing 270 more relevant documents to the solution space. Combining all systems of all methods further raises the maximum combination potential to the average precision of 0.7819 with 1725 total relevant documents retrieved.

Figure 3, as well as other optimum performance tables at various ranks, shows that potential combination exists at all ranks. In other words, the numbers in optimum performance level tables prove the existence of the combination potential by showing that combination of the retrieval results of individual systems can increase the total number of relevant documents retrieved.

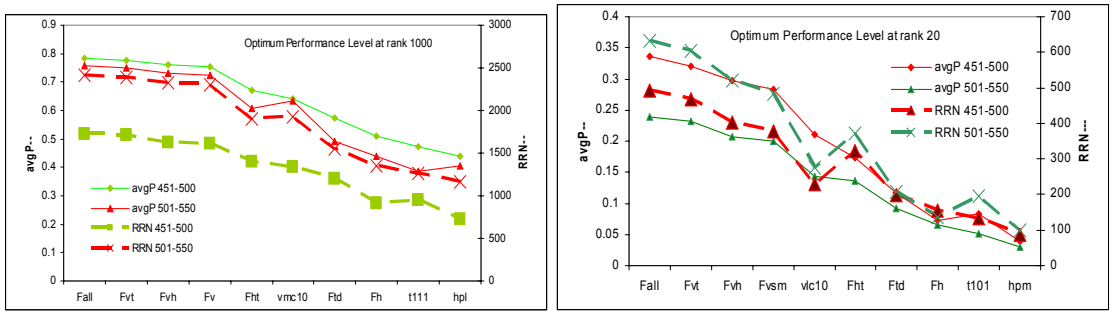

Figure 2: Optimum Performance Level at rank 1000 and 20.

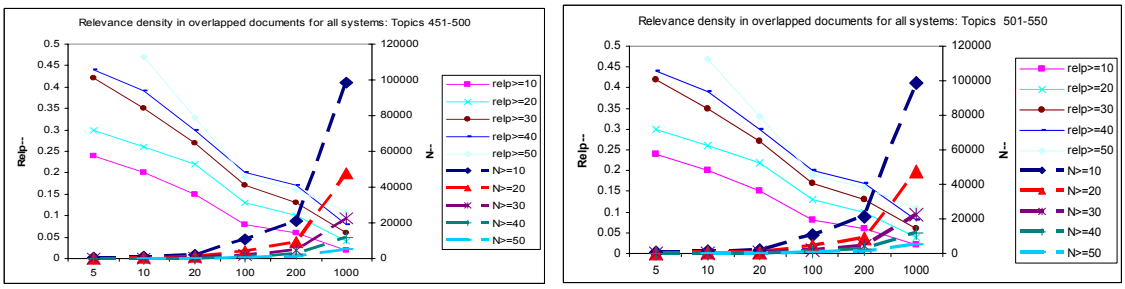

Figure 3: Relevance density in overlapped documents for all systems.

Figures 3 summarizes the overlap statistics tables by displaying the density of relevant documents at various ranks and overlap and shows the higher relevance density (i.e. proportion of relevant documents at a given overlap) not only at higher overlap but also at higher ranks. Unfortunately, the relevance density is below 50% in all but one instance (overlap >= 50 at rank 5 for topic 451-500, figure 3), which means that highly overlapped documents are more likely to be non-relevant than relevant.

In other words, overlap alone is not a good indicator of relevance because more documents are apt to be non-relevant than relevant for a given overlap despite the fact that more overlapped documents are more likely to be relevant than less overlapped documents. Table 2, which relates overlap not only to relevance but also to document ranks, shows that, in general, non-relevant documents are ranked lower than relevant documents with the same overlap in VSM and TM systems but the reverse is true for HITS systems. This peculiar pattern of overlap in HITS systems may explain why the rank-based combination formula does not behave well in HITS combination.

	Topics 451-500									*Topics 501-550*								
Rank	N1 Kdoc	p1	pV1	pH1	pt1	avgR1	avgRV1	avgRH1	avgRt1	N2 Kdoc	p2	pV2	pH2	pT2	avgR2	avgRV2	avgRH2	avgRT2
5	0.429	0.42	0.38	0.26	0.61	3.1	2.3	0.2	1.4	0.46	0.44	0.17	0.31	0.64	3.3	2.7	0.1	1.6
10	0.913	0.31	0.53	0.36	0.62	6	5	0.4	2.2	0.947	0.38	0.21	0.4	0.56	5.6	4.4	0.4	2.5
20	2	0.44	0.39	0.51	0.64	12.1	10.2	1	5.5	1.958	0.67	0.31	0.37	0.53	11.8	9.8	0.4	5.7
100	10	0.59	0.73	0.36	0.72	59.2	50.9	11.4	20.4	10.516	0.82	0.54	0.38	0.76	60	51.3	13.4	20.8
200	20	0.78	0.59	0.23	0.7	119.8	100.5	22	43.9	20.984	0.47	0.37	0.18	0.42	122.3	106.6	33.5	42.3
1000	93	0.71	0.6	0.51	0.7	624.6	530.3	32.5	238.1	98.443	0.75	0.7	0.42	0.8	609.7	512.3	34.5	211.7

Table 2: Average Ranks in overlapped documents for all systems with overlap >= 10. Column p (pV, pH, pW) shows proportion of non-relevant documents whose average ranks (of VSM, HITS, WD systems) are larger than that of relevant documents with the same overlap.

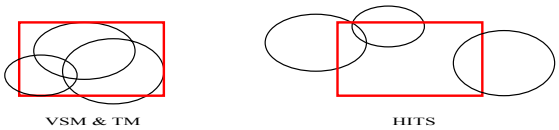

Figure 4: System solution Diagram space.

Figure 4, that display hypothetical layouts of system solution spaces (squares represent best systems), give a visual example of the combination potential. In the scenario depicted by the diagrams above, one can see that HITS systems produce more diverse solution sets than VSM or TM systems, thus resulting in a much larger combined solution space. On the other hand, the solution space of the best system in VSM and TM methods encompasses most of the combined solution space, so additional relevant documents introduced by combination with other systems are negligible.

Examination of the overlap in relevant documents confirms this hypothesis [10].

4. Conclusions

The overlap analysis revealed that the total number of relevant documents in the combined result sets of VSM, HITS, and TM systems were much more than the largest number of relevant documents retrieved by any single system. This observation suggests that the solution spaces of text-, link-, and classification-based retrieval methods are diverse enough for combination to be beneficial. It is important to note that HITS runs, despite their lower performance levels than VSM and TM runs, appeared to have the most unique contributions to the combination pool due diverse sets of relevant documents found. The high degree of unique contributions by HITS systems could be a reflection of their retrieval approach, which is distinct from VSM and TM systems with heavy reliance on text-based retrieval techniques.

The optimum performance level of the combination system is directly related to the overlap in the solution spaces of individual systems. Although the number of relevant documents retrieved has so far been used interchangeably with the solution space, it is not the only dimension of the solution space of a retrieval system. The overlap, document ranking and relevance, whose relationship we observed in the overlap statistics, are all important dimensions of the solution space [9].

One of the most important issues for combination is the optimization of the combination formula. Given less than the optimum results by individual systems, how can we combine them to bring up the ranking of the relevant documents? We have seen in the overlap analysis that, although the documents retrieved by more systems are more likely to be relevant, just the number of systems that retrieve a document is not a good indicator for relevance since highly overlapped documents were often more likely to be non-relevant than relevant. One way to compensate for this is to rely on top performing systems as was done in top system combination [1]. This, however, tends to ignore the unique contributions with its heavy emphasis on overlap. One of the most difficult challenges of top system combination, as well as combination in general, is devising a method that rewards both the overlapped and unique contributions to the combined solution space.

References

[1] Fox E. A. & Shaw J. A. (1994). Combination of multiple searches. In D. K. Harman (Ed.) *Proceedings of the Second Text Retrieval Conference* (TREC-2) (NIST Spec. Publ. 500-215 pp. 243-252).

[2] Lee J. H. (1996). Combining multiple evidence from different relevance feedback methods (*Tech. Rep. No. IR-87*). Amherst: University of Massachusetts Center for Intelligent Information Retrieval.

[3] Bartell B. T. Cottrell G. W. & Belew R. K. (1994). Automatic combination of multiple ranked retrieval systems. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.*

[4] Larkey L. & Croft W. B. (1996). Combining Classifiers in Text Categorization. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.

[5] Modha & Spangler . Clustering hypertext with applications to Web searching. *11th ACM Hypertext Conf.*

[6] Thompson. P. (1990). A combination of expert opinion approach to probabilistic information retrieval part 1: The conceptual model. *Information Processing e Management* 26(3) 371-382.

[7] J.Ferreira, A.Silva, J.Delgado, How to Improve Retrieval effectiveness on the Web, *IDAS E-commerce 2004.*

[8] Lee J. H. (1997). Analyses of multiple evidence combination. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.

[9] TREC site: http://www.ted.cmis.csiro.au /TRECWeb/access_to_data.html. <01/04/04>

[10] J. Ferreira (2004). Ways of searching Information on the Web. PhD Thesis, IST, Portugal (in Portuguese)