How the Overlap Between the Search Results of Different Retrieval Systems Correlates with Document Relevance

Anselm Spoerri

Department of Library and Information Science, School of Communication, Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901. aspoerri@scils.rutgers.edu

This paper provides an analysis of the overlap between the search results of retrieval systems that partici-pated in TREC 3, 6, 7 and 8 to provide empirical support for some of the key assump-tions guiding the design of data fusion methods. It shows that the potential relevance of a document increases exponentially as the number of search methods retrieving it increases – called the Authority Effect. It also shows that documents higher up in ranked lists and found by more sys-tems are more likely to be relevant – called the Ranking Effect. Both effects can help ex-plain why major data fusion methods perform well.

Introduction

Data fusion uses multiple sources of evidence, such as different query formulations, document representations or result sets of different search methods to infer the potential relevance of docu-ments (Callan 2000). In particular, data fusion aims to surpass the performance of individual retrieval systems by combining the results of multiple systems that search the same database. A key assumption guiding the design of data fusion methods is that documents found by multiple systems are more likely to be relevant, which will be referred to as the *Authority Effect*. However, no systematic study has been conducted to verify the validity of this assumption.

Many research papers, which describe methods for fusing the results of multiple retrieval systems, state that the Authority Effect is guiding their method design without providing evidence or references to support this design choice. It can be argued that this choice must be warranted, because fusion methods, which make explicit use of the Authority Effect, tend to out-perform methods that do not (Fox & Shaw, 1994; Lee, 1997). On the one hand, if references are provided, studies are commonly cited that use different query representa-tions to generate multiple result sets and show that a docu-ment's potential relevance increases as the number of queries retrieving it increases (Saracevic & Kantor, 1988). On the other hand, papers are cited that demonstrate how combin-ing different queries, rather than their result sets, leads to improved retrieval performance (Turtle & Croft, 1991; Belkin et al., 1993). Both types of studies provide indirect support for the Authority Effect when it comes to fusing the results sets generated by different retrieval systems. Foltz & Dumais (1992) showed that the combi-nation of result sets obtained by different methods for filtering a small collec-tion of technical reports leads to improved results; it is unclear whether these results generalize to large text databases, such as the ones used in TREC. This paper provides a sys-tematic analysis of the overlap between the search results of retrieval systems that partici-pated in TREC 3, 6, 7 and 8 to provide direct support for the Authority Effect.

The fundamental aim and hallmark of an effective retrieval method is that docu-ments higher up in its ranked list are more likely to be relevant. A further assumption guiding the

design of data fusion methods is that documents higher up in ranked lists and found by more systems are more likely to be relevant, which will be referred to as the *Ranking Effect*. The Authority Effect reflects the overlap between the result sets of the different systems. The Ranking Effect reflects the rank positions of the documents as well as how many result sets contain them.

This paper is organized as follows: first, related work is briefly discussed. Second, the methodology employed is de-scribed. Third, the analysis of the overlap between the systems participating in TREC 3, 6, 7 and 8, respectively, are presented to verify the Authority and Ranking Effects. Fourth, it is discussed how both effects provide insights into why major data fusion methods perform well. This paper also addresses how the presented results provide support for key design principles used in recent visual tools that enable users to compare the results of multiple retrieval systems. Finally, it is briefly outlined how the overlap between result sets can be used to infer the relative performance differences between the retrieval systems.

Related Work

Saracevic & Kantor (1988) used independently created Boolean queries to generate multiple result sets, and found that the greater the number of queries retrieving the same document, the greater the probability of its relevance. Foltz & Dumais (1992) found similar improvements when com-paring the result sets generated by four different filtering methods. Experiments using inference networks found that combining different query representations lead to greater retrieval effectiveness than any single representation (Turtle & Croft, 1991). Belkin et al. (1993) found that progres-sively combining different query formulations to create increasingly complex queries produced progressively im-proved performance. These results lead Belkin et al. to conclude that combining multiple pieces of evidence will enhance retrieval performance and to postulate “the more, the better” when it comes to data fusion.

Guided by the above results, major fusion methods use both voting and merging principles when combining the result sets of different retrieval methods. Fox and Shaw (1994) intro-duced a set of major methods for combining multiple results sets, including CombMNZ, Comb-SUM, CombMAX and CombMIN. When a document is found by a system, it receives a retrieval score and has a specific position in the ranked list returned by the system. Further, a document can be found by multiple systems. If a document’s retrieval scores or rank positions are normalized to a value between 0 and 1, then the sum of a document’s scores will be less or equal to the number of systems retrieving it. CombMax and CombMin are equal to the maximum or minimum score, respectively, that a document receives by the systems that find it. CombMNZ sums a document’s scores or rank positions by the different systems that find it and then mul-tiplies this sum by the number of systems that retrieve the document. CombSUM only sums a document’s scores. Lee (1997) demonstrated that CombMNZ performs best, followed by CombSUM, whereas Comb-MAX and Comb-MIN perform worst. Lee also observed that the best fusion re-sults were obtained when the systems retrieved similar sets of relevant documents and dissimilar sets of non-relevant documents. Vogt et al. verified this observation by looking at pair wise combinations of systems and they have suggested that both systems should distribute scores similarly, but not rank relevant documents similarly (Vogt & Cottrell, 1998).

Methodology

This paper analyzes the overlap between search results of retrieval systems that partici-pated in the *ad hoc track* in TREC 3, 6, 7 and 8 to provide direct support for two key assump-tions guiding data fusion. TREC provides information retrieval researchers with large document col-lections, a set of search topics and ways to compare the search results

(Voorhees & Harman, 1999). Retrieval sys-tems participating in the ad hoc task search the col-lections for each of the 50 provided topics, and then submit a ranked list of usually 1,000 documents for evaluation (50,000 docu-ments in total). For each topic, NIST pools the top 100 retrieved documents from each run. The evaluator who proposed a topic then determines the relevance of each document. The systems are evaluated based upon different measures of recall and precision. *Recall* assesses the frac-tion of relevant documents that were found by a system, while *precision* assesses the fraction of a system's retrieved documents that are relevant. The average precision for a specific topic is the mean of the precision after each rele-vant document is found. The *mean average precision* for all topics is the mean of the average precision scores.

Data fusion researchers commonly use data from the ad hoc track in TREC, because it presents an excellent environment for testing data fusion methods, since relevance judgments and result lists by many and diverse retrieval systems are readily available. Participating systems can submit multiple runs for evaluation. A run can either be *automatic* or *manual*. For the former, the query is created without human intervention based on the complete topic statement (called a *long* run) or only the title and description fields (called a *short* run).

The work presented in this paper is part of an ongoing re-search effort, whose results will be reported in several pa-pers. This paper will present an analysis of the short runs in the ad hoc track in TREC 3, 6, 7, and 8. A similar analysis is currently being conducted for the manual and long runs in TREC 3, 6, 7, and 8, and will be reported shortly.

For TREC 3, there are 19 systems that fully participated and submitted automatic runs, called category A runs. There are 24, 28 and 35 systems that submitted automatic short runs for TREC 6, 7, and 8, respectively. In terms of the total number of retrieved documents, there are 928,709 (950,000), 1,192,557 (1,200,000), 1,327,166 (1,400,000) and 1,723,929 (1,750,000) for TREC 3, 6, 7 and 8, respec-tively. The numbers in brackets represent the expected to-tals if each system retrieves 1,000 documents for all 50 topics. The actual numbers are less, indicat-ing that some system retrieve less than 1,000 documents for some topics.

In this paper, only the short run with the highest mean average precision score, called the "best" short run, is selected for each partici-pating system, because there can be a high degree of similarity between the result sets for runs of the same type and generated by the same system (Wu & Crestani, 2003). This similarity artificially boosts the Authority Effect and thus introduces a source of noise. Once the best short run for each participating system has been identified, the overlap between the result sets of the different systems is computed for each topic. Averaging across all topics, the number of documents found by a specific number of systems is computed for all documents (relevant and non-relevant) and all relevant documents.

Results

In this section, the analysis of the overlap between the best A or short runs in TREC 3, 6, 7 and 8, respectively, are presented to verify the Authority and Ranking Effects.

Authority Effect

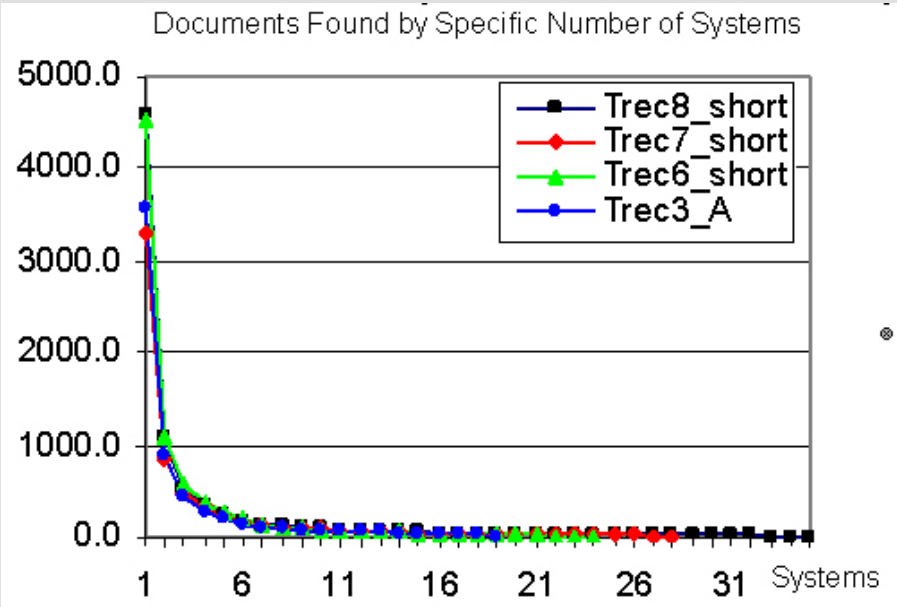
In order to test the validity of the Authority Effect, the average number of documents found by multiple systems is computed for all 50 topics. Specifically, the average number of documents only found by one system is computed, followed by the number of documents found by two systems and so on, ending with the documents retrieved by all systems that participated in the ad-hoc track and submitted short runs in TREC 3, 6, 7 and 8, respectively. Figure 1 displays the average number of docu-ments that are found by a specific

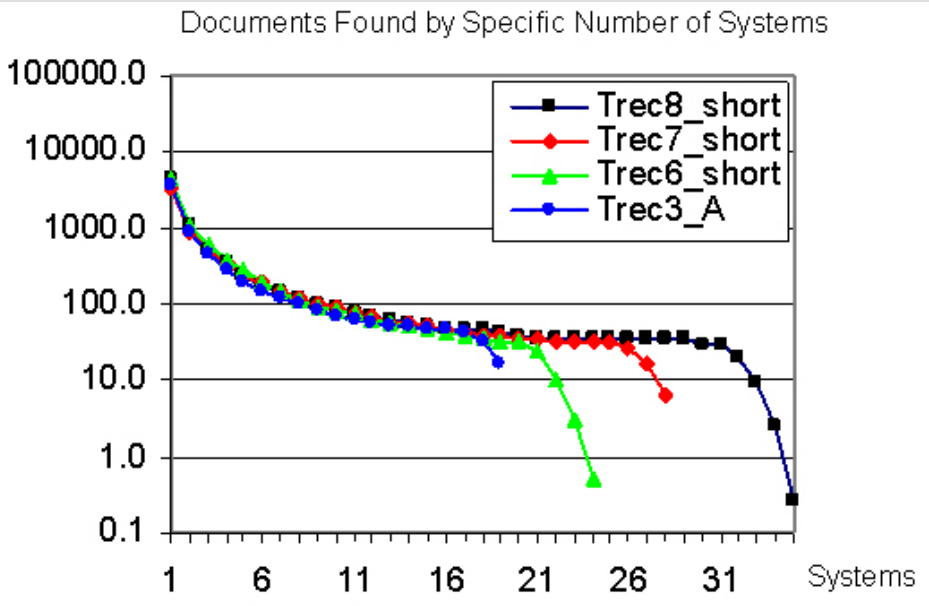
number of systems, starting with the documents retrieved by only one system and ending with those found by all systems. For TREC 8, for example, roughly 4,500 documents are found by a single system (see Figure 1 (left)). Less than 10 documents are retrieved by more than 33 systems (see Figure 1 (right)). The number of documents found by an increasing number of TREC 8 systems decreases rapidly and tends to follow a power law, which only breaks down for documents found by more than 33 systems (see Figure 2). This deviation from the power law is greatest when the worst performing systems retrieve very few relevant docu-ments and/or less than 1000 documents per topic. If the last three TREC 8 systems are not included when fitting a power law function, then the R-squared value improves from 0.75 to 0.97.

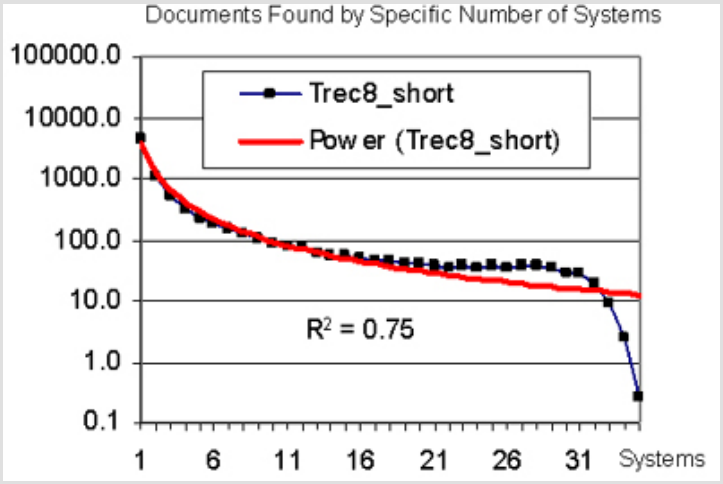
As Figure 1 illustrates, most documents are only found by a single sys-tem in TREC 3, 6, 7 and 8. The different graphs in Figure 1 have different ending points, be-cause there are 19, 24, 28 and 35 systems for TREC 3, 6, 7, and 8, respec-tively, that are being compared. For TREC 3, 6, 7 and 8, the number of documents retrieved by an increas-ing number of systems follows a power law, which only breaks down for the docu-ments found by most or all systems. If all systems are in-cluded when fitting a power law function, then the R-squared values are 0.98, 0.81, 0.96, and 0.75 for TREC 3, 6, 7, and 8, respectively. As mentioned, the deviation from the power law is greatest for the worst per-forming systems, which retrieve very few rele-vant docu-ments and/or return less than 1,000 documents per topic.

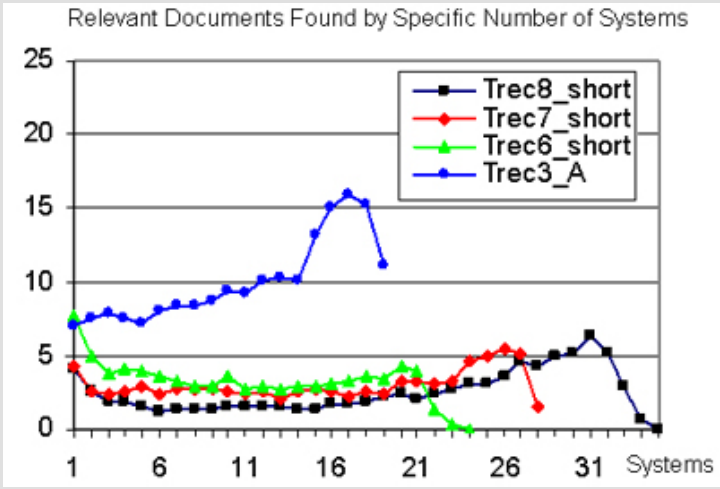
Figure 3 displays the average number of relevant documents that are retrieved by an increasing number of systems, starting with the relevant documents found by only one system and ending with those found by all systems that are being compared. For TREC 6, 7 and 8, a large number of relevant documents are found by a single system, then their average number decreases and starts to increase if more than 15 systems retrieve the same relevant document. A new peak is reached if a relevant document is found by most systems, only to decline because the worst performing systems retrieve very few relevant documents. The graph for TREC 3 stands apart from the other graphs, because the TREC 3 systems tend to retrieve twice as many relevant documents as in TREC 6, 7 or 8 (Soboroff et al. 2001). Still, the greatest number of relevant documents are found by most systems in TREC 3, only to decline because the worst performing systems find few relevant documents.

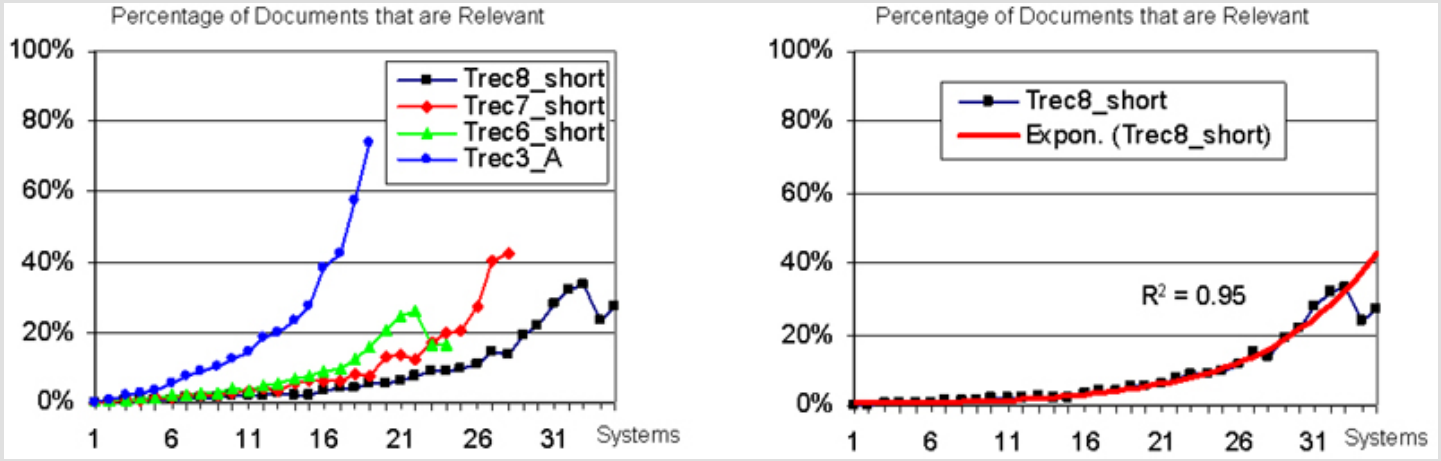
Figure 4 (left) displays the percentage of documents that are relevant and that are found by a specific number of systems for TREC 3, 6, 7, and 8, respectively. Specifically, it shows that the percentage of documents that are relevant increases exponentially as the number of systems that retrieve them increases. In Figure 4 (right) an exponential function is fitted to the graph of the percentages of TREC 8 documents that are relevant as the number of systems that retrieve them increases, resulting in a R-squared value of 0.95. The corresponding R-squared values for TREC 3, 6, and 7 are 0.90, 0.93, and 0.95, respectively. These high correlation scores provide direct support for the Authority Effect – the probability that a document is relevant increases exponentially as more of the systems that are being compared find it.


Documents Found by Specific Number of Systems
5000.0
4000.0
3000.0
2000.0
1000.0
0.0
1
6
11
16
21
26
31
Systems
Trec8_short
Trec7_short
Trec6_short
Trec3_A


Documents Found by Specific Number of Systems
100000.0
10000.0
1000.0
100.0
10.0
1.0
0.1
1
6
11
16
21
26
31
Systems
Trec8_short
Trec7_short
Trec6_short
Trec3_A


Documents Found by Specific Number of Systems
100000.0
10000.0
1000.0
100.0
10.0
1.0
0.1
1
6
11
16
21
26
31
Systems
Trec8_short
Power (Trec8_short)
R² = 0.75


Relevant Documents Found by Specific Number of Systems
25
20
15
10
5
0
1
6
11
16
21
26
31
Systems
Trec8_short
Trec7_short
Trec6_short
Trec3_A



Ranking Effect

A further goal of this paper is to investigate whether the rank position of a document combined with the information of the number of systems that retrieve it has an influence on the document's probability of being relevant. First, it is useful to pool the documents, whose rank positions lie within the same range of positions, and remove duplicate documents. For example, if the result sets of all TREC 8 systems are pooled, then the num-ber of unique documents with rank positions 1 to 50 is greater than 500, and for rank positions 951 to 1000 the number is greater than 1400 (see the thin black line in Figure 5 (left)). Figure 5 (left) displays a histogram or frequency plot of the average number of unique docu-ments as a function of their position in the ranked lists, when the 35 short runs in TREC 8 are pooled and the num-ber of documents are averaged across all 50 topics. The frequency data is col-lected using a bucket size of 50 con-secutive rank positions so that, for example, documents with rank positions between 201 to 250 are aggregated in the same bucket. If the result sets of the different systems had no documents in common, then there would be 1750 documents within each bucket, because there are 35 sys-tems and 50 documents within each range of 50 consecutive rank positions. Figure 5 (left) shows that the number of unique documents increases loga-rithmically as the rank position increases (when fitting a logarithmic function, the R-squared value is 0.99), and thus the greatest number of duplicate documents occur in the highest rank positions. A similar histogram or frequency plot can be constructed for the relevant documents. The thick blue line in Figure 5 (left) represents the average number of unique relevant documents as a function of their positions in the ranked lists for the 35 short runs in TREC 8 that are being compared. Documents with rank positions between 51 and 100 have the greatest number of unique relevant documents, and their numbers de-creases exponentially (when fitting an expo-nential function, the R-squared value is 0.97). The data displayed in Figure 5 (left) makes it possible to compute the percentage of unique documents that are relevant if a document is selected at random within a certain range of rank positions (see Figure 5 (right). As noted earlier, more than 500 unique documents and less than 50 unique relevant documents have rank positions between 1 and 50. Thus, the probability that a unique document in the top 50 is relevant is less than 10%. The percentage of unique documents that are relevant decreases stead-ily as their rank positions increase. Figure 5 (right) shows that the higher up a document is located in a ranked list (and thus the lower its rank posi-tion), the greater its prob-ability of being relevant. This result is to be expected, since retrieval systems are designed to place relevant document higher up in their result lists. Moreover, Figure 5 (right) illustrates how much can be learned about a document's potential relevance based on its rank position, if the docu-ments of all systems are pooled and duplicate documents are removed. At best, 10% of the unique documents with rank positions between 1 and 50 are relevant.

Next, the question needs to be addressed to what degree the knowledge of the number of systems that retrieved a document and of its rank position will help to identify relevant documents. If a document is found by multiple systems, then it will have multiple rank positions, which can be averaged. A low *average rank position* implies that most of the

document's rank positions are low, and thus it is placed high up in most of the result lists. Figure 6 displays the frequency plots of the average number of relevant documents found by 1, 11, 21 and 31 TREC 8 systems, respectively, as a function of their average rank positions, when the result sets of the 35 different TREC8 systems are being compared. For each specific number of sys-tems, the frequency data is aggregated using a bucket size of 50 consecutive average rank positions. It is useful to compare the average rank positions of the documents found by a specific number of systems with the resulting averages if the rank positions of a document are randomized. Such a comparison can help address the question whether or not a random process governs the position of the relevant docu-ments in the result lists. In Figure 6, the thick blue curve represents the frequency plot of the actual average rank positions, and the thin black line represents the averages of the randomized rank positions.

For the relevant documents found by one system, Figure 6 shows that the frequency plot of their actual average rank positions tends to increase toward the higher rank positions (and fitting a straight line con-firms this), whereas the plot for the randomized rank positions is "choppy" but has a horizontal tendency (fitting a straight line confirms this). For the relevant documents retrieved by 11 systems, Figure 6 shows that the frequency plot of their actual average rank positions has a peak for the range of average rank positions between 451 and 500 – in short it has a peak at 500. This fre-quency plot has a symmetrical shape that is almost identical to the plot when the rank positions are randomized. For the rele-vant documents found by 21 systems, the median of their actual average rank positions is located at 400 and the fre-quency plot has peaks close to the median, whereas the symmetrical plot of the randomized rank posi-tions has its peak at 500, and it is greater than the peak for 11 systems. Finally, for the relevant documents found by 31 systems, Figure 6 shows that the frequency plot of their actual aver-age rank positions has its peak at 100 and then steadily decreases to reach zero at 600. The symmetri-cal plot of the randomized rank positions has a higher and sharper peak at 500 than the peak for documents found by 21 systems. On the one hand, Figure 6 shows that the average rank position for the majority of relevant documents moves from the higher rank positions toward the lower ones as more systems retrieve these documents. On the other hand, the peak of the rele-vant documents with randomized rank positions remains at 500 and increases as more systems find relevant documents. Figure 6 clearly illustrates that a random process does not determine the rank positions of the relevant documents. Instead, as the number of systems retrieving the same relevant document increases, a relevant document is increasingly located toward the top of the systems' lists.

In order to be able to verify the Ranking Effect, it is necessary to compute the percentage of documents that are relevant as a function of the number of systems that find them and of their average rank positions. Figure 7 shows the frequency plots of the average number of documents that are relevant and are found by 1, 11, 21 and 31 TREC 8 systems, respectively, when the result sets of the 35 different TREC8 systems are being compared. For each specific number of systems, the frequency data is aggregated using a range of 50 con-secutive average rank positions. For the documents found by one system, Figure 7 shows that the percentage of documents that are relevant is practically zero regardless of the documents' average rank positions. For the documents retrieved by 11 systems, the frequency plot of the percent-age of documents that are relevant reaches its maximum of 10% when their average rank positions lie within the range of 151 and 200 – in short the peak is located at 200. For the documents found by 21 systems, a maximum of 30% is reached at 150. Finally, for the documents retrieved by 31 systems, Figure 7 shows that almost 60% of the documents with an average rank position in the top 50 are relevant. The percentage of documents that are relevant decreases exponentially as the average rank position increases, reaching zero for positions greater than 550.

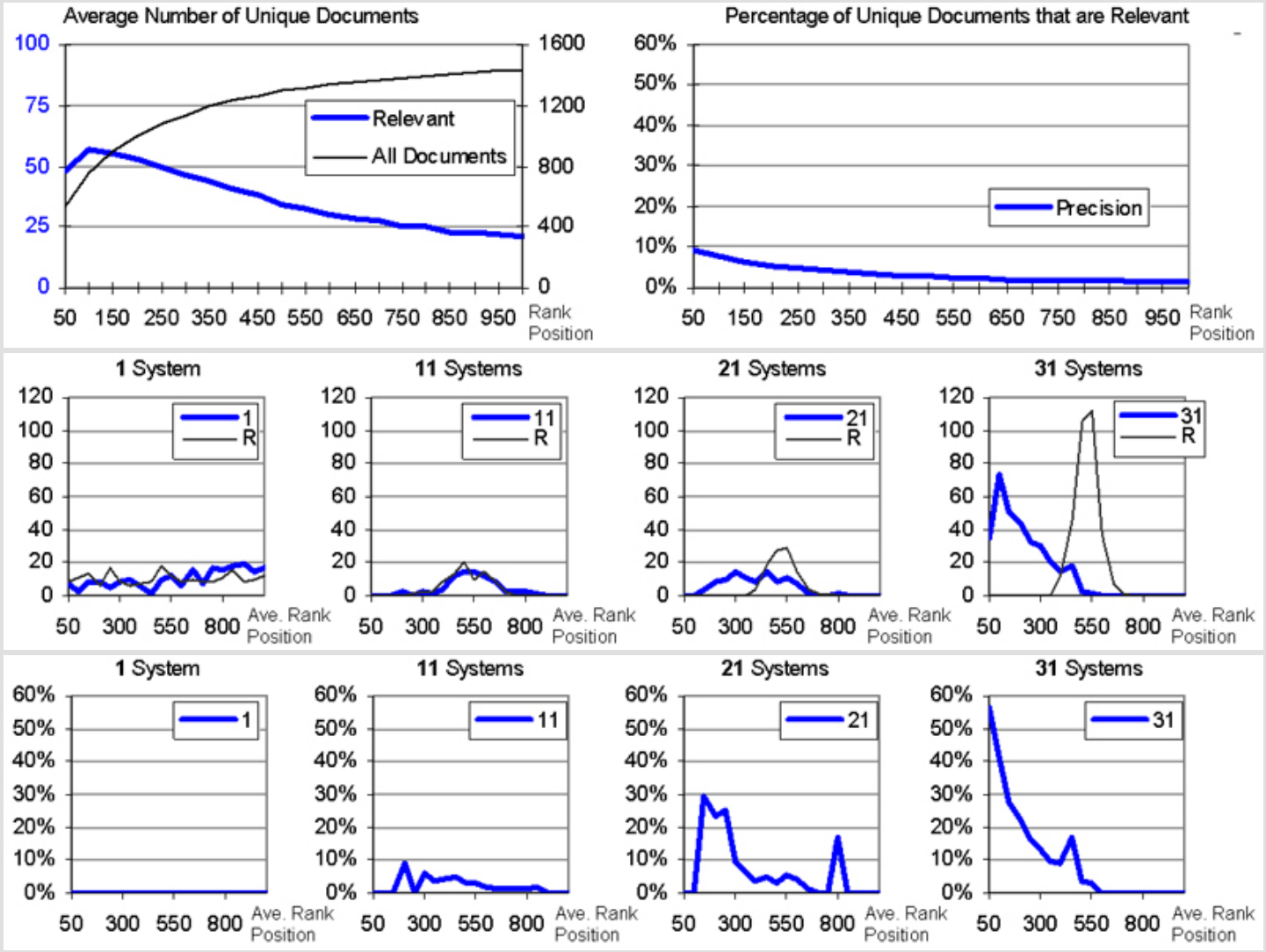


Figure 7 dem-onstrates that the probability that a document is relevant greatly increases as more systems find it and the higher up it is placed in the multiple ranked lists. These results provide direct support for the Ranking Effect. The results shown in Figures 5, 6 and 7 are based on the TREC 8 data, and the TREC 3, 6 and 7 data produce very similar results – the more systems that retrieve a document and the lower its average rank position, the more likely the document is relevant.

Discussion and Implications

As mentioned, Lee (1997) demonstrated that the data fusion method CombMNZ performs best, followed by CombSUM, whereas Comb-MAX and Comb-MIN perform worst. The results presented in this paper provide insights into why the data fusion methods CombMNZ and CombSUM perform well. Both CombSUM and CombMNZ make use of the Ranking Effect, because they sum the normalized scores or rank positions – the higher up a document in multiple lists, the greater the sum. This summing operation also incorporates the Authority Effect, because the more systems that find a document, the more scores or rank positions can be added. However, this summing operation, and thus CombSUM, does not sufficiently take advantage of the Authority Effect. The CombMNZ multiples CombSUM by the number of systems that find a document, which makes the Authority Effect dominant. On the one hand, being high up in the ranked lists, thus having a strong Ranking Effect, ampli-fies the Authority Effect. On the other hand, a weak Authority Effect can mute a strong Ranking Effect. The CombMAX and CombMIN only consider the information provided by the rank positions, and they do not reward documents that are retrieved by multiple systems. The CombMIN penalizes documents that are found by multiple systems, because the probability that all of its rank positions are high is low. Further, Figure 7 can help explain why CombMNZ ensures that relevant documents are placed toward the top of the list that is created when multiple result sets are fused. CombMNZ stretches each frequency plot by the number of systems that retrieve the documents. These stretched histograms are then overlapped with their right corners coinciding. This CombMNZ transformation tends to move many of the relevant documents to the very left and thus toward the top of the fused list.

This paper has shown that the Authority Effect impacts data fusion in a major way and it will be outlined below that it also provides insights into the effectiveness of the systems being fused (Spoerri 2005). The Authority Effect has also been referred to as the "Chorus" or "Popularity" Effect (Vogt & Cottrell, 1998; Aslam & Savell, 2003). These latter two terms seem to suggest that the overlap between the result lists is a minor event and they do not do justice to its power. The decision taken in this paper to use only the run with the highest mean average precision that are submitted by each group participating in TREC, instead of com-paring all the runs, helped "sharpen the signal," thereby making the importance of the Authority Effect more readily visible.

The analysis presented in this paper is motivated in part by current research in information visualization. In particular, the MetaCrystal visual toolset has been developed to enable users to visually compare the result sets of multiple search engines (Spoerri 2004). A key design principle used in MetaCrystal is to map the docu-ments in such way that users can use the distance from the display's center as a visual cue of a document's po-tential relevance. Thus, MetaCrystal maps the documents found by multiple systems and with high average rank positions toward the center of the display. The work previously cited by Saracevic & Kantor (1988), Foltz & Dumais (1992) and Belkin et al. (1995) provide some support for the document mapping employed in the MetaCrystal visualization. The results presented in this paper provide definite support: MetaCrystal's tools visualize the overlap between search results of different retrieval methods in ways that fully exploit the Authority and Ranking Effects. In particular, MetaCrystal's Category View displays the number of documents found by different numbers of search methods. Usually, subsets of up to five different retrieval systems are compared in a single Category View. Spoerri (2005) shows that the Authority and Ranking Effects are present when smaller numbers of retrieval systems are being compared than the 19, 24, 28 and 35 systems for TREC 3, 6, 7, and 8, respec-tively, that were compared in this paper. The question arises what other type of insights the MetaCrystal visualization enables in addition to which documents are more likely to be relevant. Spoerri (2005) shows that the percent-ages of documents found by a specific number of retrieval systems changes in a systematic way as the quality of the systems being compared decreases. Specifically, the greatest per-centage of documents is found by all five systems and then shifts toward the documents retrieved by a single system as the mean average precision of the five systems being compared decreases. This systematic change in the overlap structure can be used to infer the effectiveness of the methods being fused without the need for relevance judgments. In par-ticular, it can be shown that the percentage of documents found by a single system is negatively correlated with the mean average precision of the systems being compared (Spoerri, 2005).

Conclusions

This paper has provided a systematic analysis of the overlap between search results of the retrieval systems that partici-pated in the ad hoc track in TREC 3, 6, 7 and 8 to provide empirical support for two key assump-tions guiding data fusion. It showed that the number of documents found by an increasing number of systems follows a power law. More importantly, it demonstrated that a document's probability of being relevant increases expo-nentially as the number of search methods retrieving it increases – called the *Authority Effect* – and thereby providing direct support for the primary assumption guiding the design of data fusion methods. This paper showed that the placement of the relevant documents in ranked lists is not a random process. Instead, as the number of systems retrieving the same relevant document increases, a relevant document is increasingly located toward the top of the systems' lists. This paper dem-onstrated that a document's probability of being relevant increases greatly as more systems find it and the higher up it is placed in the multiple ranked lists – called the *Ranking Effect* – and thereby providing direct support for another key assumption guiding the design of data fusion methods. The presented results also provided insights into why the major data fusion methods, such as CombMNZ and CombSUM, perform well. Specifically, CombMNZ tends to perform best because it emphasizes the Authority Effect while exploiting the Ranking Effect.

Acknowledgements

The author would like to thank Nick Belkin, Paul Kantor and Tefko Saracevic for their help and feedback. This research has been supported by a Rutgers Research Council Grant. The TREC data used in the research reported in this paper has been provided by NIST and can be downloaded at www.nist.org.

References

Aslam, J. & Savell, R. (2003). On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2003)*.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management, 31*(3), 431-448.

Callan. J. (2000). Distributed information retrieval. In Croft W.B. (Ed.), *Advances in Information Retrieval* (pp. 127-150). Kluwer Academic Publishers.

Foltz, P. & Dumais, S. (1992). Personalized information delivery: An analysis of information-filtering methods. *Communications of the ACM, 35*(12), 51-60.

Fox, E. & Shaw, J. (1994). Combination of Multiple Searches. In *2nd Annual Text Retrieval Conference (TREC-2)*, Gaithersburg, MD: National Institute of Standards and Technology.

Lee, J. H. (1997). Analyses of Multiple Evidence Combination. In *Proceedings of the 20th Intl. Conference. on Research and Development in Information Retrieval (SIGIR'97)* (pp. 267–276).

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches and overlap. *Journal of the American Society for Information*

Science. 39(3) 197-216.

Spoerri, A. (2004). Visual Search Editor for Composing Meta Searches. In *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology (ASIST 2004)*.

Spoerri, A. (2005). *Using the Structure of Overlap Between Search Results to Rank Retrieval Systems Without Relevance Judgments*. Unpublished manuscript.

Soboroff, I., Nicholas, C. and Cahan, P. (2001) Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR 2001)* 66-73.

Turtle, H., Croft, B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*.

Vogt, C. & Cottrell, G. (1998). Predicting the performance of linearly combined IR systems. In *Proceedings of the 21st ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 190–196).

Voorhees, E. & Harman, D. (1994) Overview of the Eighth Text REtrieval Conference (TREC-3). In *The Eighth Text REtrieval Conference (TREC-8)*. Gaithersburg, MD: U.S. Government Printing Office.

Voorhees, E. & Harman, D. (1997) Overview of the Eighth Text REtrieval Conference (TREC-6). In *The Eighth Text REtrieval Conference (TREC-8)*. Gaithersburg, MD: U.S. Government Printing Office.

Voorhees, E. & Harman, D. (1998) Overview of the Eighth Text REtrieval Conference (TREC-7). In *The Eighth Text REtrieval Conference (TREC-8)*. Gaithersburg, MD: U.S. Government Printing Office.

Voorhees, E. & Harman, D. (1999) Overview of the Eighth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*. Gaithersburg, MD: U.S. Government Printing Office.

Wu, S. & Crestani, F. (2003). Methods for Ranking Information Retrieval Systems Without Relevance Judgements. In *Proceedings of ACM SAC'03* (pp. 811- 816).