Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1059–1070
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Using the structure of overlap between search results to rank retrieval systems without relevance judgments

Anselm Spoerri *

School of Communication, Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901, USA

Received 5 June 2006; received in revised form 10 September 2006; accepted 11 September 2006
Available online 15 December 2006

Abstract

This paper addresses the problem of how to rank retrieval systems without the need for human relevance judgments, which are very resource intensive to obtain. Using TREC 3, 6, 7 and 8 data, it is shown how the overlap structure between the search results of multiple systems can be used to infer relative performance differences. In particular, the overlap structures for random groupings of five systems are computed, so that each system is selected an equal number of times. It is shown that the average percentage of a system's documents that are only found by it and no other systems is strongly and negatively correlated with its retrieval performance effectiveness, such as its mean average precision or precision at 1000. The presented method uses the degree of consensus or agreement a retrieval system can generate to infer its quality. This paper also addresses the question of how many documents in a ranked list need to be examined to be able to rank the systems. It is shown that the overlap structure of the top 50 documents can be used to rank the systems, often producing the best results. The presented method significantly improves upon previous attempts to rank retrieval systems without the need for human relevance judgments. This "structure of overlap" method can be of value to communities that need to identify the best experts or rank them, but do not have the resources to evaluate the experts' recommendations, since it does not require knowledge about the domain being searched or the information being requested.

Keywords: Ranking retrieval systems; Data fusion; Meta search; Result overlap analysis

1. Introduction

Test collections play an important role in the field of information retrieval (IR) because they enable IR researchers to evaluate and compare different retrieval methods in a controlled setting. The construction of these document collections is expensive and time consuming, because of the need to determine the relevance of a large number of documents with respect to a set of search topics. The TREC workshops address this problem by pooling the top 100 documents retrieved by each participating system and then an evaluator determines the relevance of each document in the pool (Voorhees & Harman, 1999). Zobel (1998) has shown that

* Tel.: +1 732 932 7500 8211; fax: +1 732 932 2644.
E-mail address: aspoerri@scils.rutgers.edu

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.09.009

this pooling method leads to reliable results in terms of determining the effectiveness of retrieval systems and their respective rankings, but the relevance determination process is still very resource intensive.

This paper addresses the problem of how to rank retrieval systems in the absence of or without the need for a set of human relevance judgments. The proposed method is easy to compute and utilizes the phenomenon that retrieval systems tend to retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents (Lee, 1997). Specifically, it computes the structure of overlap between the search results of random groupings of five retrieval systems. It will be shown that the percentage of a system's documents that are only found by it and no other system increases as the system retrieval quality decreases, since a poorly performing system tends to find fewer relevant documents, causing its results to overlap less with other systems. This result is then used to demonstrate that the average percentage of a system's documents that are only found by it and no other systems is strongly and negatively correlated with its retrieval effectiveness. Thus, the relative performance differences between the systems can be inferred based on the percentage of a system's documents that are not found by any of the other systems. An analogy may help to explain the solution presented in this paper: there is a group of professed experts and the task is to rank them based on their long lists of recommended documents to a set of information requests without knowing which documents represent relevant recommendations. First, one selects a subset of five experts, compares their list of recommended documents and then counts how many of each expert's documents were only suggested by him or her and nobody else. In order to obtain a representative count, each expert is included an equal number of times in random subsets of five experts. Next, one computes and averages the percentages of an expert's documents that only s/he recommended. Which expert should be "trusted" the most and which one the least? This paper proposes that the expert with the lowest percentage of documents that are not recommended by any other expert should command the greatest trust, whereas the expert with the highest percentage of documents only recommended by him or her should be trusted the least. In a way, the proposed method defines and operationalizes "quality of expertise" in terms of the degree of consensus or agreement an expert can generate. Since the proposed method does not require knowledge or expertise about the domain being searched or the content of the information being requested, it could be of value to communities that need to identify the best experts or rank them, but do not have the resources to evaluate the experts' recommendations. This paper will show that only the top 50 documents per information request are needed to infer system rankings of high accuracy in the context of retrieval systems searching the same database.

This paper is organized as follows: first, related work is discussed. Second, the methodology employed is described. Third, the overlap structures between systems of decreasing quality are computed and it is shown that there is a systematic change in the structure of overlap as the retrieval performance of the systems decreases. Fourth, results using data from TREC 3, 6, 7 and 8 are presented that show how the overlap structure between the search results of random groupings of five systems is strongly correlated in a negative way with the relative performance differences between the systems. Finally, the presented work and future research are discussed.

2. Related work

Prior research relevant or related to the work presented falls into two categories: (1) research examining how the overlap between search results is related to the potential relevance of documents; (2) methods proposed for ranking retrieval systems in the absence of relevance judgments.

2.1. Identification of relevant documents

Saracevic and Kantor (1988) found that the greater the number of independently created Boolean queries that retrieve the same document, the greater the probability of its relevance. Foltz and Dumais (1992) found similar results when comparing the results obtained by four different methods designed to filter a technical reports database. Belkin, Kantor, Fox, and Shaw (1995) combined different query formulations to create increasingly complex queries that produced a progressively improved retrieval performance. Guided by these results, major data fusion methods use both voting and merging strategies to combine the result sets of different retrieval systems (Fox & Shaw, 1994). It has been observed that data fusion leads to improved retrieval

performance if there is a greater overlap of relevant documents than of non-relevant documents (Lee, 1997; McCabe, Chowdhury, Grossman, & Frieder, 1999; Vogt & Cottrell, 1998). Specifically, the number of relevant documents found by all systems divided by the number of all relevant documents found needs to be greater than the same ratio for the non-relevant documents (Lee, 1997). Spoerri (2005) has conducted a systematic analysis of the overlap between the results of retrieval systems that participated in TREC 3, 6, 7 and 8. The analysis showed that the potential relevance of a document increases exponentially as the number of systems retrieving it increases – called the Authority Effect. Further, this analysis showed that the higher the positions of a document in the ranked list that contain it and the greater the number of systems that retrieve it, the greater probability of the document being relevant – called the Ranking Effect. These two effects suggest that the overlap between search results provides a rich source that can be mined further. A key contribution of this paper is that it shows how the overlap structure – the percentage of a system's retrieved documents that are also found by a specific number of other systems – can be used to infer the relative performance differences between retrieval systems, significantly improving upon existing methods.

2.2. Ranking retrieval systems

Only a few methods have been proposed to rank retrieval systems without the need for human relevance judgments, because this is a difficult problem, since there does not exist a proven theory for determining expertise without explicit relevance or correctness judgments. A contribution of this paper is that it suggests a working hypothesis (and potential theory) for how to rank retrieval systems in the absence of human relevance judgments.

Soboroff, Nicholas, and Cahan (2001) address the problem of how to rank retrieval systems without the need for human relevance judgments by generating a set of pseudo-relevance judgments by randomly selecting and declaring some documents from the pool of top 100 documents as relevant. This set of pseudo-relevance judgments (instead of a set of human relevance judgments) is then used to determine the effectiveness of the retrieval systems. For each topic, the average number of relevant documents in the pool has to be known to decide how many pseudo-relevant documents to select in the pool of top 100 documents. These pseudo-relevance judgments are used to estimate the performance of a system, using data from the ad hoc track in TREC 3, 5, 6, 7 and 8. The Kendall's tau correlation with the actual TREC rankings range from 0.37 in TREC-7 to 0.49 in TREC-5, and top-performing systems are ranked together with the poorly performing systems. Soboroff et al. found that the greatest improvement came from retaining duplicate documents in the pools, which is consistent with the Authority Effect. Aslam and Savell (2003) devised a simple way to measure the similarity between two retrieval systems by computing the ratio of the number of documents in their intersection and union. This measure produced results that were highly correlated with the method using pseudo-relevance judgments, similarly ranking the best systems with poor performers. Aslam and Savell hypothesize that this is caused by a "tyranny of the masses" effect and that these two related methods are assessing the systems based on "popularity" instead of "performance." The analysis by Spoerri (2005) suggests that the "popularity" effect is caused by considering all the runs submitted by a retrieval system, instead of only selecting one run per system, which would help to "sharpen the signal" and make the Authority Effect more dominant.

Wu and Crestani (2001, 2003) have developed two related "reference counting" methods to rank retrieval systems without relevance judgments. Both methods assign a score to each retrieved document based on the number of systems that have found it as well as taking into account the rank positions of its duplicates or the scores assigned by the systems. In essence, the "reference counting" methods attempt to leverage the Authority and Ranking Effects, but they do so only in partial ways, since ad-hoc formulas are used to compute a document's score that not fully reflect the structural properties of these two effects (Spoerri, 2005). Using TREC 5, 7, 9 and 10 data, the initial method by Wu and Crestani (2001) just counted how many of a system's top documents were also found by a random selection of two to nine other systems. This simple "reference count" was used to divide the systems into three categories: good, fair and poor using heuristic thresholds. The TREC data was used to classify the systems into three categories. If the reference counting approach classified a system as "good" but its TREC category was "poor", then this was counted as a "big mistake." Wu and Crestani ran 50,000 trials using TREC 2001 data, resulting in a mistake rate of 30.4%, and a big mistake rate of 5.2%.

In their second method, Wu and Crestani (2003) developed multiple ways for computing the ''reference count'' to rank retrieval systems and they compared these variations with the ranking method by Soboroff et al., using TREC 3, 5, 6, 7, and 10 data. In Wu and Crestani's implementation of the pseudo-relevance method (PR), 10% of the top 100 documents are declared as relevant, whereas the original method requires the number of relevant documents to be drawn from a normal distribution with a mean and standard deviation derived from the actual number of relevant documents per topic in the pool. Further, Wu and Crestani made the distinction between an ''original'' document and its duplicates in all other lists, called the ''reference'' documents, when computing a document's score. The basic method (Basic) computes for each ''original'' document how many systems retrieved it. The first variation (V1) assigns different weights to ''reference'' documents based on their ranking positions. The second variation (V2) assigns different weights to the ''original'' document based on its ranking position. The third variation (V3) consists of assigning different weights to the ''reference'' documents as well as the ''original'' document based on their respective rank positions. The fourth variation (V4) uses the ''original'' document's normalized score instead of its ranking position and the ''reference'' documents' ranking positions to assign the weights. Wu and Crestani also vary the number of systems used to compute the ''reference count'' (ranging from 3 to 20 systems). For each number of systems, they construct 20 randomized runs for the 50 search topics, and then average the results. The Spearman rank correlation is used to compare their system rankings with the official TREC rankings. If the latter rankings are based on the mean average precision, then the variation V4, which uses the ''original'' document's normalized score and the ''reference'' documents' ranking positions to assign the weights, performed the best among the five methods. However, none of them performed as well as the pseudo-relevance method (PR), which performed best for TREC 3 with a Spearman rank correlation of 0.63 and worst for TREC 7 with a correlation value of 0.41.

Wu and Crestani (2003) show that the similarity between the multiple runs submitted by the same retrieval system affect the ranking process. If only one run per systems is selected, then the variation (V3), which takes into account of both the ranking positions of the ''original'' and ''reference'' documents, outperforms PR by 45.5% on average for 3–9 systems, whereas PR outperforms V3 by 6.5% for 10 and more systems. Further, the accuracy of the ranks assigned to the best performing systems is considerably improved for all methods. This suggests that the similarity between the multiple runs by the same system greatly affects the top end of the rankings. Wu and Crestani find that the ''reference counting'' approach performs best if the inferred system rankings are compared with the TREC rankings based on the systems' precision@100 measures (followed by R-Precision, and worst for average mean precision). Further, the Spearman rank correlations of the ''reference counting'' and pseudo-relevance judgments rankings with respect to the official TREC rankings steadily increase as more systems are considered, starting to reach a plateau when random selections of 10 systems are used.

3. Methodology

The TREC workshops provide IR researchers with large document collections, a set of search topics and relevance judgments (Voorhees & Harman, 1994, 1997, 1998, 1999). This makes it possible to compare and analyze the effectiveness of different retrieval methods in a controlled setting. Retrieval systems participating in the ad hoc task search the collections for each of the 50 provided topics, and then submit a ranked list of usually 1000 documents per topic for evaluation (and 50,000 documents in total). The ad hoc task is similar to how a person would search a static and known collection of documents, but the search requests can vary (Voorhees & Harman, 1999). As mentioned, for each topic, the top 100 retrieved documents from each run are pooled and then an evaluator determines the relevance of each document in the pool. Each system participating in TREC 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). In this paper, the short runs in TREC 3, 6, 7 and 8 are used, because a greater number of systems submitted short runs (Voorhees & Harman, 1994, 1997, 1998, 1999). However, only one short run (i.e., the one with the highest ''mean average precision'', where this performance measure will be explained below) is selected for each participating 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 boosts the degree of overlap artificially and thus introduces a source of noise. There are 18 (19), 24, 25 (28) and 35 best short runs for TREC 3, 6, 7 and 8, respectively, that are ranked in this paper. The numbers in brackets indicate the total number of different systems, and some systems were not included in this study because they submitted significantly less than 50,000 documents in total. When several systems are compared, the overlap between their result sets is computed for each topic. Averaging over all topics, the number of documents found by a specific number of systems is computed. In particular, it is informative to compute the average percentage of a system's documents that are also found by a specific number of multiple systems. For example, if system A retrieves 1000 documents for a topic and this result set is compared with the results of four other systems, then we compute the percentage of A's documents that are found by all five systems, the percentage of A's documents found by four systems and so on, ending with the percentage of A's documents that are only found by the system A itself.

The major measures used to evaluate the performance of the systems participating in TREC are: (1) the *mean average precision* for all topics, which is equal to the mean of the averages of the precision values after a relevant document is retrieved for each topic; (2) the *R-precision*, which is the average of the precision values after R documents have been retrieved, where R is the number of relevant documents for each topic; (3) the *precision at 1000*, which is the average of precision values at 1000 documents for each topic. These different performance measures tend to be highly correlated. In this paper, the mean average precision and precision at 1000 are used to create the official TREC rankings that are then used to evaluate the effectiveness of the new ranking method presented.

4. Structure of overlap

This paper proposes that the overlap structure between multiple search results can be used to infer the quality of the systems being compared without the need to know which documents are relevant. To better understand why the structure of overlap can provide this type of insight, it is instructive to compute the overlap between five "consecutive" systems in TREC 8 that have been ordered based on their mean average precision, starting with the top five systems (1–5) and ending with bottom five runs (31–35). Fig. 1 (left) shows that on average more than 30% of the documents found by a top five system are also retrieved by all of the top five systems, whereas the bottom five systems find almost no such documents. On average, almost 20% of the documents are retrieved by a single top five system, whereas roughly 70% are found by a single bottom five system. Similarly, Fig. 1 (middle) shows that on average more than 60% of the relevant documents found by a top five system are also retrieved by all of the top five systems, whereas very few relevant documents are found by all of the bottom five systems. On average, less than 10% of the relevant documents are retrieved by a single top five system, whereas more than 40% are found by a single bottom five system. Finally, Fig. 1 (right) displays the percentage of documents that are relevant and shows that this percentage increases exponentially as the number of systems finding the same document increases.

Fig. 1 shows that the percentages of a system's documents that are found by a specific number of systems changes in a systematic way as the quality of the five systems being compared decreases. Specifically, the

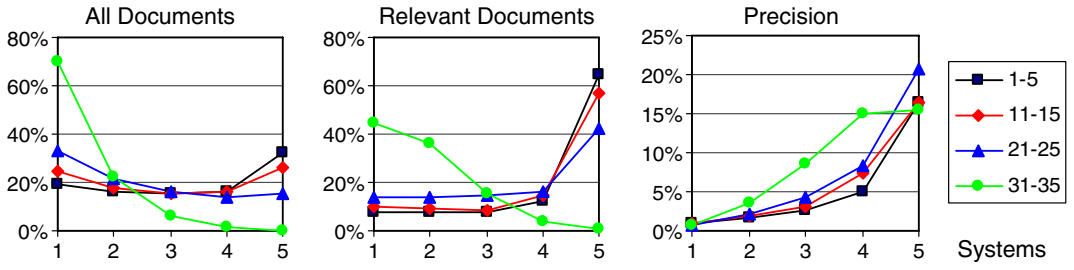

Fig. 1. Average percentage of (relevant) documents retrieved by a system that are also found by a specific number of other systems, when five "consecutive" systems in TREC 8 are compared, starting with the top 1–5 and ending with the bottom 31–35 systems, where the systems are ordered based on their mean average precision.

greatest percentage of documents is found by all top five systems and then shifts toward the documents retrieved by a single bottom five system as the mean average precision of the systems being compared decreases. This change in percentages is most pronounced for the documents found by a single system or by all systems. This systematic change in the structure of overlap can be used to infer the effectiveness of the methods being fused without the need for relevance judgments. It will be shown that the percentages of a system's documents not found by other systems (Single%) as well as the difference between the percentages of documents found by a single system and all five systems (Single% – AllFive%) are highly and negatively correlated with the mean average precision and precision at 1000 scores of the systems.

It has been suggested that different systems searching the same database tend to retrieve similar sets of relevant documents but return different sets of non-relevant documents (Lee, 1997). Fig. 1 supports this suggestion, because it shows that the degree of overlap decreases as the quality of the systems, and thus the number of relevant documents found, decreases. Fig. 1 shows that the systems ranked 1–25 have a large set of relevant documents in common, which only diminishes when the worst performing systems, which find only few relevant documents, are compared.

5. Ranking retrieval systems

The question arises whether the systematic change in the overlap structure, which occurs when five "consecutive" systems of increasingly lower retrieval quality are compared, still occurs for random groupings of five systems. First, the overlap structure needs to be computed for a sufficient number of random groupings. Second, each system needs to be selected an equal number of times. If there are N systems to be ranked, then the randomization process can be constrained so that each system appears five times in the N random groupings of five systems. Next, for each random grouping and each of the 50 topics, the percentage of documents found by a specific number of systems needs to be computed for each system. These percentage values are then averaged over the 50 topics. Finally, for each system, the N random groupings produce five percentage values (see Fig. 2) that can be averaged as well (see Fig. 3).

It would seem that the overlap structure should vary greatly if a specific system is compared with different sets of systems, especially if these other systems are mostly top five systems in one comparison, and mostly bottom five systems in another one. In the former, it is expected that a large percentage of a system's documents are found by all five systems, whereas in the latter few documents are expected to be retrieved by all systems. Thus, poorly performing systems can greatly affect the percentage of documents found by all five systems. However, the percentage of documents retrieved by a single system is more robust, because a high percentage of such documents are not relevant (see Fig. 1 (right)). Fig. 2 (left) displays a scatter plot of the mean average precision scores for the 35 systems in TREC 8 versus all of the five percentage values of a system's documents that are only found by the system itself (*Single%*). This figure shows that there is a strong linear and negative correlation between Single% and mean average precision. Fig. 2 (middle) confirms the expectation that the percentage of a system's documents that are found by all five systems (*AllFive%*) is not such a

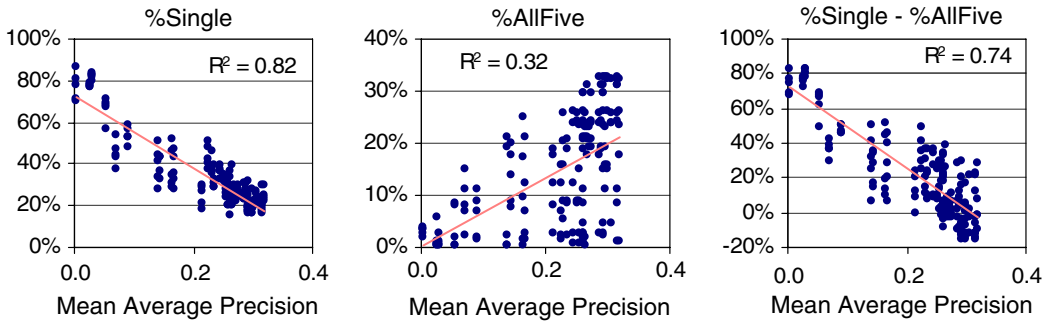

Fig. 2. Scatter plots of the TREC 8 systems' mean average precision versus all their percentages of documents found by only one system (Single%), by all five systems (AllFive%) and Single% minus AllFive%, respectively, for 35 random groupings of five systems.

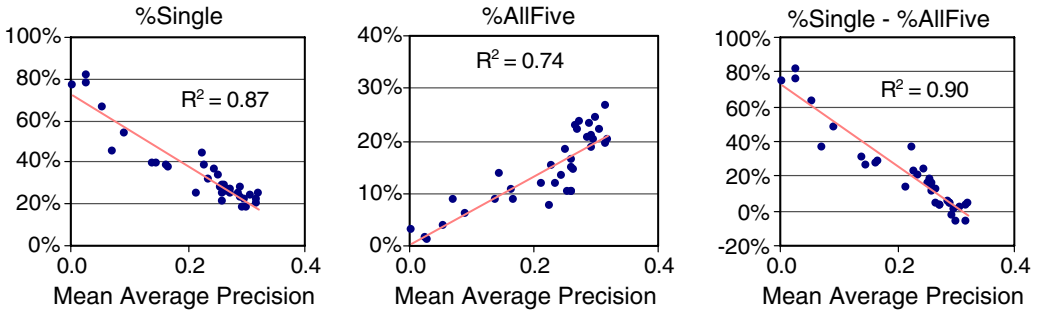

Fig. 3. Scatter plots of the mean average precisions scores for the 35 TREC 8 systems versus their *average* percentages of documents found by only a single system (Single%), all five systems (AllFive%) and the difference between Single% and AllFive% for 35 random groupings of five systems.

robust indicator of system quality, because the AllFive% values are very scattered, although an increasing percentage score is weakly correlated with a system's mean average precision. Fig. 2 (right) displays a scatter plot of the mean average precision scores versus the difference between Single% and AllFive%, and it shows that there is a strong linear and negative correlation, although Single% minus AllFive% is slightly more scattered than Single%.

Next, the multiple percentage values for each system can be averaged. Fig. 3 displays the scatter plots of these average percentages versus the mean average precisions of the 35 TREC 8 systems. The R-squared values for the linear correlations for Single%, AllFive% and Single% minus AllFive% are 0.87, 0.74 and 0.90, respectively. The R-squared value for the average of the AllFive% values is greatly improved with respect to its value for all AllFive% values (see Fig. 2 (middle)).

These average percentage values can be used to rank the 35 systems in TREC 8. Fig. 4 displays a scatter plot of the official TREC 8 rankings based on the systems' mean average precision scores versus the rankings based on the average Single%, AllFive% and Single% minus AllFive%, respectively. A more appropriate measure for comparing the different rankings is the Spearman rank correlation measure, which is 0.89, 0.88 and 0.95 for Single%, AllFive% and Single% minus AllFive%, respectively. Fig. 4 shows that for the TREC 8 systems the difference between Single% and AllFive% produces higher Spearman rank and R-squared linear correlation values than if Single% and AllFive% are considering separately. The Single% minus AllFive% values seem to be both sensitive to the number of non-relevant documents, which tend to be retrieved primarily by a single system, and the number of relevant documents, which tend to mainly found by multiple systems. If a system retrieves many relevant documents, then a majority of these documents tend to be found by other systems, reducing Single% and potentially boosting AllFive%. If a systems retrieves few relevant documents, then the great majority of its documents will be found only by the system in question, significantly boosting Single% and AllFive% becomes insignificant.

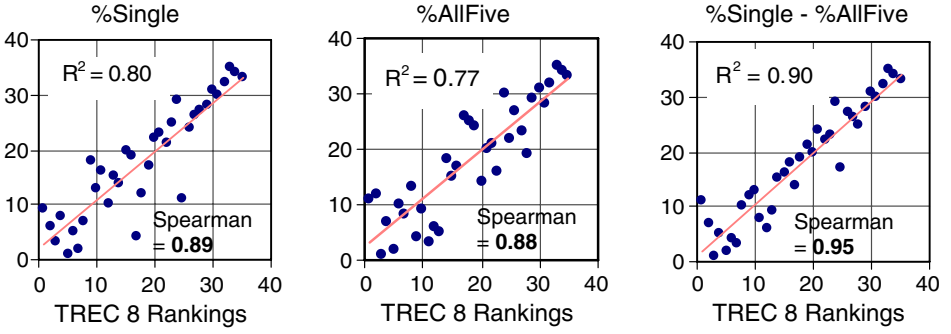

Fig. 4. Spearman rank correlations if the 35 TREC 8 systems are ranked based on their average Single%, AllFive%, Single% minus AllFive%, respectively, versus the official TREC 8 system rankings based on their mean average precision scores.

In the analysis so far, all of the 1000 documents retrieved per topic by a system are used to rank the systems. The question arises whether even better results could be obtained if only, for example, the top 50 documents are used to compute the overlap structure. The Ranking Effect suggests that it should be possible to infer the relative retrieval differences using the top 50 or 100 documents, because relevant documents tend to be located higher up in a ranked list, especially if they are found by multiple systems (Spoerri, 2005). For TREC 8, 7, 6 and 3, Fig. 5 (left column) displays the R-squared values if the linear correlation between the average Single% and Single% minus AllFive% values is calculated with respect to the systems' mean average precision (MAP),

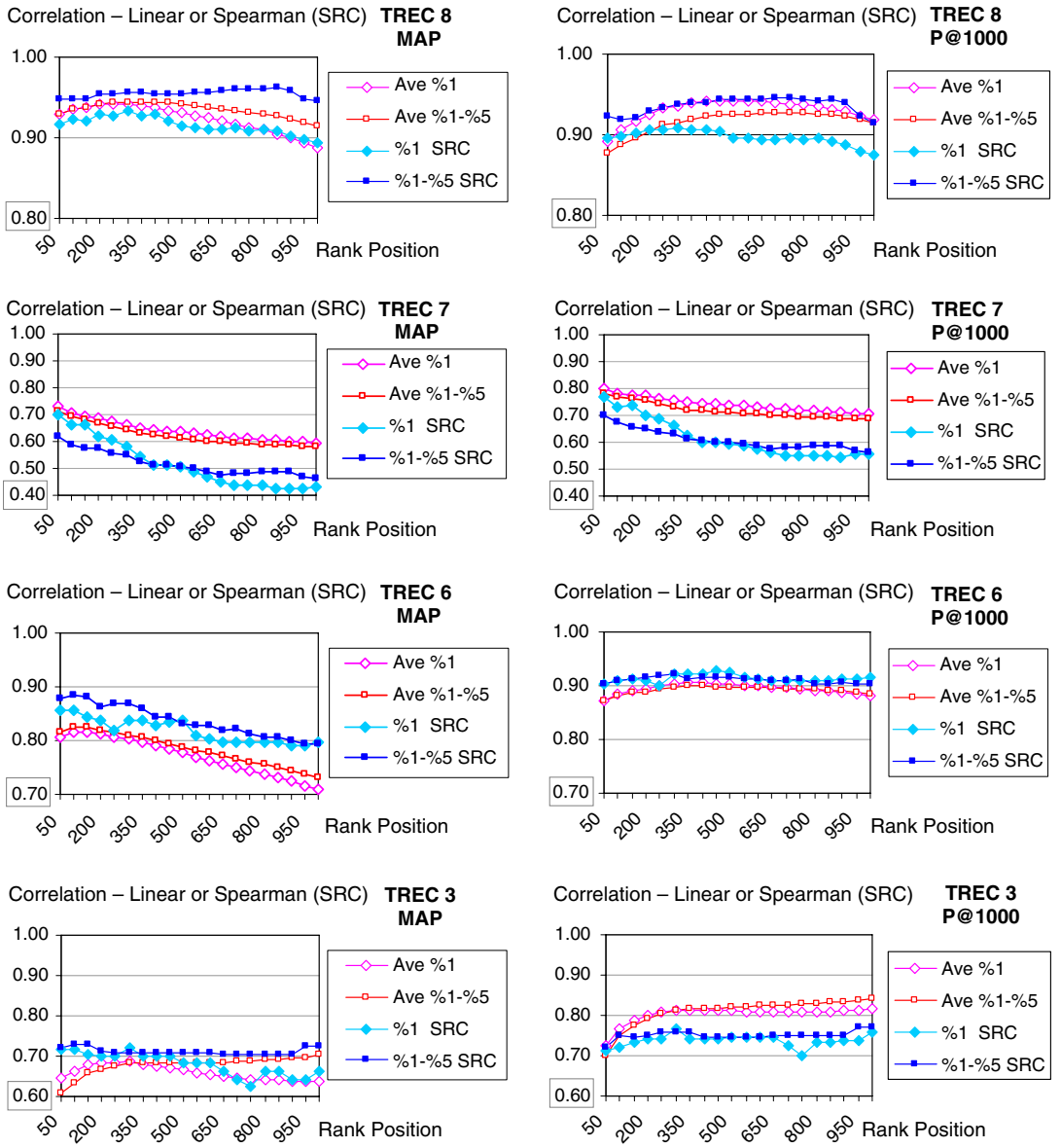

Fig. 5. Displays the R-squared values if the linear correlation (outlined shapes) between the average Single% (diamonds shapes) and Single% minus AllFive% (square shapes and denoted by %1 and %5, respectively) are calculated with respect to the systems' mean average precision (MAP) or precision at 1000 scores, as the overlap structure is computed using an increasing number of documents in the ranked lists. The Spearman rank correlation scores (solid shapes) are also displayed for Single% and Single% minus AllFive%, respectively, as documents increasingly lower in the ranked lists are included in the overlap computation. Note that the displays for TREC 8, 7, 6 and 3 do not all use the same minimum value on the vertical correlation axis to make the differences between the different graphs better visible.

Table 1
summarizes the Spearman rank correlation scores

MAP	Single%		Single%–AllFive%	
	Max	Top 50	Max	Top 50
TREC 3	0.72	0.72	0.73	0.72
TREC 6	0.86	0.86	0.88	0.88
TREC 7	0.70	0.70	0.62	0.62
TREC 8	0.93	0.92	0.96	0.95
P@1000	Single%		Single%–AllFive%	
	Max	Top 50	Max	Top 50
TREC 3	0.76	0.71	0.77	0.72
TREC 6	0.93	0.90	0.92	0.90
TREC 7	0.77	0.77	0.70	0.70
TREC 8	0.91	0.90	0.95	0.92

and the overlap structure is computed using an increasing number of documents in the ranked lists. The right column shows the correlations with respect to the precision at 1000 scores (P@1000). Fig. 5 also displays the Spearman rank correlation scores for Single% and Single% minus AllFive%, respectively, as documents increasingly lower in the ranked lists are included in the overlap computation (note that the displays in Fig. 5 do not all use the same minimum value on their vertical correlation axes to make the differences between the different graphs better visible). For TREC 3, Fig. 5 shows that the rankings based on the top 50 documents provide the best results for MAP, whereas for P@1000, examining all 1000 documents produces the best rankings. For TREC 6, rankings based on the overlap between the top 100 documents produce the best or close to the best ranking results for MAP and P@1000, respectively. For TREC 7, the top 50 documents produce the best rankings for both MAP and P@1000. For TREC 8, the rankings based on the top 50 documents are very good, and the correlation values for all rank positions are very high. Table 1 displays the maximum Spearman rank correlation and the resulting correlation if only the top 50 documents are examined to compute the overlap structure for TREC 3, 6, 7 and 8 data. In summary, Fig. 5 and Table 1 show that if only the top 50 documents are compared, the resulting system rankings produce correlation values with respect to the official TREC rankings that are close to, if not equal, to maximum correlation values that are obtained if more than the top 50 documents are considered.

6. Discussion and future research

The "structure of overlap" method presented in this paper is simple and effective. As Table 1 shows, the system rankings produced by this method have a high Spearman rank correlation with the official TREC rankings. If the percentage of a system's documents that are only found by the system itself (Single%) is used to rank the systems, then the Spearman rank correlation values range from 0.70 to 0.93 when the mean average precision (MAP) scores are used to create the TREC rankings, and from 0.71 to 0.93 if precision at 1000 (P@1000) is used. If the difference between the percentages of a system's document that are found by a single system and all five systems (Single% minus AllFive%) is used to ranks the systems, then the Spearman rank correlation values range from 0.62 to 0.96 for MAP, and from 0.70 to 0.95 for R@1000. The results for Single% and Single% minus AllFive% tend to be very similar, as are the results for MAP and P@1000. These Spearman rank correlation scores are significantly better than the rank correlation scores obtained by Wu and Crestani (2003), who compared several "reference counting" methods and the pseudo-relevance method by Soboroff et al. (2001), where the latter performed best for TREC 3 with a Spearman rank correlation of 0.63 and worst for TREC 7 with a correlation value of 0.41. Further, these methods attain their best results when 10 or more systems are used and many randomized runs need to be conducted. The "structure of overlap" method requires only N random groupings of five systems to effectively rank N systems. Future research will also investigate the minimal and optimal number of systems that need to be compared to rank retrieval systems without the need for relevance judgments. Future research will also investigate how well the retrieval systems can be ranked and their relative performance differences inferred if all the systems are compared at once, instead of multiple random

subsets of systems. If all systems are compared at once, the issue may arise that similarities between the algorithms used by the different systems may affect the overlap structure and it will be investigated if and how retrieval systems using similar text retrieval and analysis methods could be detected.

Both the ranking methods developed by Soboroff et al. (2001) and Wu and Crestani (2003) produce poor results for the best performing systems because they are ranked together with the poorly performing systems. As Fig. 4 shows, the "structure of overlap" method ranks the top-performing systems correctly, and performs equally well for the systems in the medium range and the poorly performing systems. There are, however, ranking errors for the top 10 systems with respect to each other, since there are small quality differences between the top ten systems and it is not surprising that some the top five systems are ranked in the top six to ten positions and visa versa.

The question arises whether the "*Overlap Effect*", which refers to the systematic difference in the overlap structure between retrieval systems of varying quality, can also be observed at the level of the individual topics, and not just when the results for the 50 topics are averaged. Ongoing research suggests that the Overlap Effect is present for most individual topics, and that the overlap structure for an individual topic can be used to rank the systems. Thus, the Overlap Effect, which is observed when the results for all the 50 topics are combined, is not an artifact of averaging. Now, some topics produce better rankings and/or have more documents that are relevant. Future research will investigatie how to detect the topics that produce robust rankings, while taking into account that a retrieval system does not perform equally well for all topics.

The "structure of overlap" method is straightforward to compute and only the top 50 documents found by a system have to be examined to produce a correlation value with respect to the official TREC rankings that is close to, if not equal, to maximum correlation value that can be obtained if a greater number of documents are considered. In particular, the overlap structure for N random groupings of five systems needs to be computed. For each random grouping, the degree of overlap between 250 documents has to be determined for each of the 50 topics. For TREC 8, for example, less than half million documents need to be examined. The above discussion regarding the Overlap Effect at the level of the individual topics suggests that not all topics need to be considered. For example, a subset of 25 topics that have varying numbers of relevant documents would be sufficient, reducing the number of documents to be examined in half.

As mentioned, it is critical that only one run by each participating system is included, because otherwise the similarity between the search results is artificially increased. In this paper, the run with the highest mean average precision was selected for each system. Instead of selecting the best run, we could have selected any of the multiple runs submitted by each participating system to compute the overlap structure. Future research will investigate how to detect a high degree of similarity between search results so that systems that tend to use very similar retrieval methods can be identified.

The question arises of how the data and system rankings computed by the "structure of overlap" method can be used to improve existing data fusion methods. For example, fusion methods have been proposed that use linear combination models or apply weights when combining multiple result lists (Vogt & Cottrell, 1999; Wu & Crestani, 2001). The percentage of a system's documents that are found by no other systems can be used to specify a weight to be applied when combining the different result sets, since this percentage value tends to be strongly correlated in a linear and negative way with a system's mean average precision (see Fig. 3).

Both Soboroff et al. (2001) and Wu and Crestani (2003) have suggested that their methods could be employed in the context of the World Wide Web, where the databases used by different Web search engines are tremendously large and change continuously. However, their methods and the approach presented in this paper have been tested with retrieval systems searching the same database, whereas it has been estimated that Web search engines only index 20% of the Internet and their databases overlap to varying degrees (Lawrence & Giles, 1999). There could be many relevant web pages that are only found by a single Web search engine. Thus, the Web search engines may tend to retrieve dissimilar sets of relevant documents and dissimilar sets of non-relevant documents. Soboroff et al. noted that their pseudo-relevance method performed better if duplicates were not removed. In the case of overlapping databases, there will be fewer duplicates. Similarly, the "reference count" approach will have fewer documents that receive high scores, which are critical for inferring relative performance differences between the systems. Future research will address how the "structure of overlap" approach can be modified so that it can be applied in the context of multiple systems searching overlapping, but not identical databases.

The "structure of overlap" method can complement and facilitate the current TREC "pooling" method used to identify relevant documents. It can help human evaluators identify highly performing systems, whose documents are more likely to be relevant, especially if these documents have high rank positions and are found by multiple systems. Thus, it can help the evaluators identify relevant documents more quickly. The "structure of overlap" method can also support Interactive Searching and Judging (ISJ), which involves a small of group of evaluators who use a retrieval system to search a TREC database for the documents that are related to the TREC search topics (Cormack, Palmer, & Clarke, 1998; Sanderson & Joho, 2004). The evaluators then judge the relevance of the documents contained in a result list, starting with the top documents and stopping when the frequency of relevant documents encountered becomes such that continuing appears unproductive. Using TREC 6 data, Cormack et al. (1998) showed that the ISJ approach can produce a set of found relevant documents that can be used to evaluate the effectiveness of retrieval systems, but requires fewer documents to be judged then the TREC "pooling" method. The "structure of overlap" method can be used to help identify effective retrieval systems that can be used by the experts in the ISJ process.

7. Conclusions

Using TREC 3, 6, 7 and 8 data, this paper showed how the overlap structure between the search results of random groupings five systems can be used to rank retrieval systems without the need for human relevance judgments. First, it showed that there is a systematic change in the overlap structure when five "consecutive" systems of increasingly lower retrieval quality are compared. Second, it was demonstrated that the average percentage of a system's documents found only by it and no other system is strongly and negatively correlated with both its mean average precision and precision at 1000, as is the difference between the percentages of documents found by a single system and all five systems being compared. This result was used to develop the "structure of overlap" method, which uses the degree of consensus or agreement a retrieval system can generate to infer its quality and thus the relative performance differences between the systems. Third, this paper showed that if only the top 50 documents are compared, the resulting system rankings produce correlation values with respect to the official TREC rankings that are close to, if not equal, to maximum correlation values that are obtained if a greater number of documents are considered. Fourth, the presented method significantly improves upon previous attempts to rank retrieval systems without relevance judgments. The "structure of overlap" method can be of value to communities that need to identify the best experts or rank them, but do not have the resources to evaluate the experts' recommendations, since the presented method does not require knowledge about the domain being searched or the content of the information being requested.

Acknowledgments

The author would like to thank Nick Belkin and Paul Kantor as well as the reviewers for their valuable feedback. The TREC data used in the research reported in this paper has been provided by NIST and can be downloaded at http://trec.nist.gov/. This research was supported by a Rutgers research Council grant.

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 international ACM conference on research and development in information retrieval (SIGIR-2003)* (pp. 361–362).

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining evidence of multiple query representations for information retrieval. *Information Processing and Management, 31*(3), 431–448.

Cormack, G. V., Palmer, C. R., & Clarke, C. L. A. (1998). Efficient construction of large test collections. In: *Proceedings of the 21st annual international ACM conference on research and development in information retrieval (SIGIR 1998)* (pp. 282–289).

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)* (pp. 243–252), NIST, Gaithersburg, MD.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature, 400*, 107–109.

Lee, J. H. (1997). Analyses of multiple evidence combination. In: *Proceedings of the 20th annual international ACM conference on research and development in information retrieval (SIGIR'97)* (pp. 267–276).

McCabe, M. C., Chowdhury, A., Grossman, D., & Frieder, O. (1999). A unified environment for fusion of information retrieval approaches. In: *Proceedings of the 8th annual international ACM conference on information and knowledge management (CIKM 1999)* (pp. 330–334).

Sanderson, M., & Joho, H. (2004). Forming test collections with no system pooling. In: *Proceedings of the 26th annual international ACM conference on research and development in information retrieval (SIGIR 2004)* (pp. 33–40).

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. (2005). How the overlap between search results correlates with relevance. In: *Proceedings of the 68th annual meeting of the American Society for Information Science and Technology (ASIST 2005)*.

Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In: *Proceedings of the 24th annual international ACM conference on research and development in information retrieval (SIGIR 2001)* (pp. 66–73).

Vogt, C., & Cottrell, G. (1998). Predicting the performance of linearly combined IR systems. In: *Proceedings of the 21st annual international ACM conference on research and development in information retrieval (SIGIR 1998)* (pp. 190–196).

Vogt, C., & Cottrell, G. (1999). Fusion via a linear combination of scores. *Information Retrieval, 1*(3), 151–173.

Voorhees, E., & Harman, D. (1994). Overview of the third text retrieval conference (TREC-3). *The third text retrieval conference (TREC-3)*, Gaithersburg, MD, USA, 1994. U.S. Government Printing Office, Washington.

Voorhees, E., & Harman, D. (1997). Overview of the sixth text retrieval conference (TREC-6). *The sixth text retrieval conference (TREC-6)*, Gaithersburg, MD, USA, 1997. U.S. Government Printing Office, Washington.

Voorhees, E., & Harman, D. (1998). Overview of the seventh text retrieval conference (TREC-7). *The seventh text retrieval conference (TREC-7)*, Gaithersburg, MD, USA, 1998. U.S. Government Printing Office, Washington.

Voorhees, E., & Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). *The eighth text retrieval conference (TREC-8)*, Gaithersburg, MD, USA, 1999. U.S. Government Printing Office, Washington.

Wu, S., & Crestani, F. (2001). Data fusion with estimated weights. In: *Proceedings of the 11th annual international ACM conference on information and knowledge management (CIKM 2001)* (pp. 648–651).

Wu, S., & Crestani, F. (2003). Methods for ranking information retrieval systems without relevance judgements. In: *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC'03)* (pp. 811–816).

Zobel, J. (1998). How reliable are the results of large-scale retrieval experiments? In: *Proceedings of the 21st annual international ACM conference on research and development in information retrieval (SIGIR 1998)* (pp. 307–314).

Anselm Spoerri is an Assistant Professor at SCILS, Rutgers University. He was a researcher at AT&T Bell Labs after completing his Ph.D. research at MIT, where he developed InfoCrystal. He holds a B.Sc. in Mathematics from the University of London, England, a M.Sc. in Computational Vision and a Ph.D. in Information Visualization, both from MIT.