Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1044–1058
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Examining the Authority and Ranking Effects as the result list depth used in data fusion is varied

Anselm Spoerri *

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

Received 8 July 2006; received in revised form 5 September 2006; accepted 7 September 2006
Available online 28 November 2006

Abstract

The Authority and Ranking Effects play a key role in data fusion. The former refers to the fact that the potential relevance of a document increases exponentially as the number of systems retrieving it increases and the latter to the phenomena that documents higher up in ranked lists and found by more systems are more likely to be relevant. Data fusion methods commonly use all the documents returned by the different retrieval systems being compared. Yet, as documents further down in the result lists are considered, a document's probability of being relevant decreases significantly and a major source of noise is introduced. This paper presents a systematic examination of the Authority and Ranking Effects as the number of documents in the result lists, called the *list depth*, is varied. Using TREC 3, 7, 8, 12 and 13 data, it is shown that the Authority and Ranking Effects are present at all list depths. However, if the systems in the same TREC track retrieve a large number of relevant documents, then the Ranking Effect only begins to emerge as more systems have found the same document and/or the list depth increases. It is also shown that the Authority and Ranking Effects are not an artifact of how the TREC test collections have been constructed.

Keywords: Data fusion; Meta search; Authority; Ranking effects

1. Introduction

Data fusion aims to improve the performance of individual retrieval systems by combining the results of multiple systems, usually searching the same database. It uses multiple sources of evidence to decide how best to combine the different result sets and produce a new ordering of the documents that moves relevant ones – the "signal" – further toward the top of the fused result list and non-relevant documents – the "noise" – toward the bottom of the fused list (Callan, 2000). A key piece of information used to help "separate the signal from the noise" is the number of systems that found a document, which Spoerri (2005) calls the *Authority*

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

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.09.005

Effect, and the assumption is made that the greater the number of systems that found a document, the greater its probability of being relevant (Fox & Shaw, 1994). The Authority Effect reflects the phenomenon that retrieval systems searching the same database tend to retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents (Lee, 1997). A further key piece of data used in data fusion is the average rank position of a document in the multiple lists that contain it, which Spoerri (2005) refers to as the *Ranking Effect*. Major data fusion methods assume that documents placed higher up in multiple lists and found by more systems are more likely to be relevant (Fox & Shaw, 1994). Spoerri (2005) has examined the correlation between a document's probability of being relevant and the number of systems that find it and its positions in the lists that contain it. Specifically, this research provided *direct* support for the Authority and Ranking Effects, whereas previous work in data fusion assumed their validity and the success of methods that leveraged these effects provided at best indirect support.

A central challenge for data fusion methods is to identify relationships between the different result lists that make it possible to detect all or many of relevant documents (high recall) and move relevant documents closer toward the top of the fused result list (high precision). The work described in this paper is part of a research program that investigates what can be learned by only comparing and analyzing the result lists of different retrieval systems searching the same database without the need to know anything about the specialized methods used to produce the result lists or to employ other analytical tools. In a certain sense, the approach taken is analogous to the "micro or technical analysis" used by foreign currency traders who analyze the changes in price and sales volume of currencies to infer whether to buy or sell specific currencies. These "micro analysis" traders assume that all the "market knowledge and wisdom" is reflected in the currency price charts, whereas "macro analysis" traders use macro-economic fundamentals and specialized data to develop sophisticated models to infer how to trade foreign currencies. By analogy, each retrieval system employs a "macro analysis" approach that uses specialized data and sophisticated analytics to decide which documents are most likely to be relevant to produce a ranked result list that represents its best estimate of relevance. The data fusion methods studied in this paper employ a "micro analysis" approach because they analyze multiple ranked lists and use easily observable information, such as a document's different rank positions and the number of lists that contain it, to identify potentially relevant documents. A contribution of this paper is that it shows that such a "micro or technical analysis" approach can be effective and that the Authority and Ranking Effects reflect robust relationships between a document's probability of being relevant and the observed overlap between different result sets.

Many and some of the most effective data fusion methods use merging and voting methods, which is equivalent to computing and using the overlap structure between the different result lists (Fox & Shaw, 1994; Lee, 1997). In particular, documents that are found by multiple methods and placed high up in the respective lists are promoted. Thus, relevant documents founds by many systems receive higher scores than relevant documents only found by one system, which has the effect that the fused list is less likely to have a greater recall than the best systems being compared. However, the fused list tends to have increased precision, because the merging and voting methods leverage the Authority and Ranking Effects, which cause relevant documents to be promoted toward the top of the fused list. Thus, a major advantage of merging and voting data fusion methods is their ability to improve precision. Further, these fusion methods guide users toward documents that may be most relevant or useful to them, because the top documents in the fused list tend to be found high up in the result lists of many systems (i.e., they are highly recommended by many "experts"). Using TREC 12 Robust data, this paper will show that documents judged as "highly relevant" tend to be found by more systems than documents judged as simply "relevant". Data fusion offers complimentary benefits with respect to single retrieval engines, because it provides users with feedback about which documents are most likely to be relevant, because such documents tend be found by many systems and are placed toward the top of the lists that contain them.

It is common for data fusion methods to use all the documents returned by the different retrieval systems. On the one hand, it can be argued that "more is better" and the fact that a document is found by many systems, although placed by some toward the bottom of their result lists, provides very useful information to identify relevant documents. On the other hand, retrieval methods aim to place the potentially relevant documents toward the top of their result lists and as documents further down in the result lists are considered, a document's probability of being relevant decreases significantly and a major source of noise is introduced.

Thus, it is useful to study the consequences of varying the number of documents examined in the result lists, called the *list depth*, on the Authority and Ranking Effects, and thus the ability to "separate the signal from the noise" and to identify potentially relevant documents.

Using TREC 3, 7, 8, 12 and 13 data, this paper presents a systematic analysis of the validity of the Authority and Ranking Effects as the list depth is varied. This paper is organized as follows: first, related work is briefly discussed. Second, the methodology employed is described. Third, the distribution of relevant documents is examined at different list levels to gain insight into how this may affect the Authority and Ranking Effects. Fourth, a document's probability of being relevant is computed as function of the list depth used and the number of systems that have found it to investigate the Authority Effect. Fifth, a document's probability of being relevant is computed as a function of the list depth used, the number of systems that have found it and its average rank position in the lists containing it to examine the Ranking Effect. Sixth, it is shown that the Authority and Ranking Effects are not an artifact of how the TREC test collections have been constructed, where only top 100 documents are pooled and examined to identify relevant documents. Seventh, it is shown that documents judged as "highly relevant" are found by more systems than documents judged as simply "relevant." Finally, it is discussed how the effectiveness of major data fusion methods is affected by varying the list depth as well as the question is addressed whether the Authority and Ranking Effects may be affected, since the systems participating in TREC may adopt the most successful solutions of previous TREC years.

2. Related work

As briefly noted above, Spoerri (2005) conducted an analysis of the overlap between the search results of the retrieval systems that participated in the short tracks in TREC 3, 6, 7 and 8 to provide empirical support for the Authority and Ranking Effects. This analysis showed that the number of documents found by an increasing number of systems declines and follows a power law. More importantly, it was demonstrated that a document's probability of being relevant increases exponentially as the number of systems retrieving it increases, thereby providing direct support for the Authority Effect. Next, it was shown 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. Further, it was demonstrated 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, thereby providing direct support for Ranking Effect.

In terms of previous research that provides indirect support for the Authority and Ranking Effects, Saracevic and 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 and Dumais (1992) found similar improvements when comparing 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, Kantor, Fox, and Shaw (1995) found that progressively combining different query formulations to create increasingly complex queries produced progressively improved 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.

Fox and Shaw (1994) introduced a set of major methods for combining multiple results sets, such as CombMNZ and CombSUM, where Lee (1997) demonstrated that CombMNZ performs best, followed by CombSUM, in terms of data fusion effectiveness. When a document is found by a system, it 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 rank positions are normalized to a score between 0 and 1 (the higher up in the result list, the greater the score), then the sum of a document's scores will be less or equal to the number of systems retrieving it. CombSUM only sums a document's scores. CombMNZ sums a document's scores by the different systems that find it and then multiplies this sum by the number of systems that retrieve the document. CombMNZ and CombSUM exploit to varying degrees the Authority and Ranking Effects. Both of them make use of the Ranking Effect, because they sum the normalized 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 are added. The Authority Effect is more dominant for CombMNZ than for CombSUM, since CombMNZ multiples CombSUM by the number of systems that find a document.

3. Methodology

The TREC workshops provide IR researchers with large text collections, a set of search topics and relevance judgments to enable the comparison and analysis of the effectiveness of different retrieval methods in a controlled setting (Craswell & Hawking, 2004; Voorhees & Harman, 1994, 1998, 1999; Voorhees, 2003, 2004). This paper uses the ranked lists returned by the retrieval systems that participated in the *ad hoc track* in TREC 3, 7 and 8, the *robust track* in TREC 12 and the *web track* (*distillation task*) in TREC 13 to examine the Authority and Ranking Effects as the list depth is varied. Firstly, the TREC 3, 7, 8, 12 and 13 years (1994, 1998, 1999, 2003, 2004, respectively) were chosen, because they represent a diverse subset of all the TREC years, where the retrieval systems participating in the selected tracks search different document collections for the different 50, 75 and 100 provided topics in the ad hoc, web (distillation task) and robust tracks, respectively. Secondly, the systems in the chosen TREC years submit a ranked list of usually 1000 documents per topic for evaluation, which provides data fusion methods with a large number of documents to fuse as well as makes it possible to study the effects of varying the list depth. In order to identify the relevant documents, the top 100 retrieved documents (or the top 125 documents for topics 51–100 in the TREC 12 Robust track) are pooled from each submitted result list for each topic and then an evaluator determines the relevance of each document in the pool. The systems are evaluated based upon different measures of recall and precision. *Recall* assesses the fraction 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 values after each relevant document is found. The *mean average precision* for all topics is the mean of the average precision scores.

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, 7 and 8 are used, because a greater number of systems submitted short runs (Voorhees & Harman, 1994, 1998, 1999). The TREC 12 Robust track is chosen, since the 100 topics used in this track consist of 50 topics selected from the ad hoc tracks in TREC 6–8 that proved especially difficult for most retrieval systems and 50 new topics that were selected with the expectation to be difficult as well (Voorhees, 2003). Firstly, this makes it possible to examine the Authority and Ranking Effects for poorly performing topics. Secondly, the systems were able to explicitly train on the difficult topics chosen from TREC 6–8, but systems were not allowed to use the available relevance judgments when producing the result lists that were submitted. This makes it possible to examine to what degree the training of the systems tends to produce similar result lists. For the TREC 13 Web track, the 75 distillation topics are chosen, since for these topics the systems needed to find more than one relevant web page and tended to return 1000 documents per topics (Craswell & Hawking, 2004; Voorhees, 2004). In summary, the selected ad hoc, robust and web TREC tracks and years make it possible to investigate the Authority and Ranking Effects in diverse and difficult settings.

As mentioned, each system can submit multiple runs for evaluation in a track, but only the run with the highest mean average precision score, called the "best" run, is used for each system in this study, 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. There are 18 (19), 25 (28), 35, 16 (17) and 11 (17) best runs for TREC 3, 7, 8, 12 and 13 respectively, that are analyzed in this paper. The numbers in brackets indicate the total number of different systems in a track, and some systems were not included in this study because they submitted significantly less than 1000 documents on average per topic. Once the best run for each system has been identified (although any single run submitted by a system could be used), the overlap between the result sets of the different systems in the same track in a specific TREC year is computed for each topic. Next, 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, respectively.

4. Results

First, the question will be addressed how the distribution of relevant documents changes as the list depth is varied. Second, data will be presented that shows that the Authority Effect is present at all list depths. Third, it will be demonstrated that the Ranking Effect is present at all list depths, but if the systems in a TREC year retrieve a large number of relevant documents, then the Ranking Effect only begins to emerge as more systems have found the same document and/or the list depth increases. Fourth, it will be shown that an increasing majority of new documents, which are found at a specific list depth level, are only found by a single system and very few new documents are found by more than three or four systems for list depths greater than 100 and approaching 1000 documents. Fifth, it will be shown that documents judged as "highly relevant" are found by more systems than documents judged as "relevant".

4.1. Distribution of relevant documents

For each TREC year studied, the average number of unique relevant documents found per topic is computed if all of the 1000 documents returned per topic are considered. This number is 188, 85, 87, 72, 33 and 21 for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, respectively (see Fig. 1(left)). TREC 3 has more than double the average number of relevant documents per topic than TREC 7, 8 or 12a (topics 1–50); six and nine times the number of relevant documents than TREC 12b (topics 51–100) and TREC 13, respectively. This offers the opportunity to explore how the number of relevant documents affects the Authority and Ranking Effects; it will be shown that the average number of relevant documents per topic influences the strength of these two effects. Next, the percentage of the total number of relevant documents, which has been found by all systems if only the top N documents are considered, can be plotted, where the percentages for all topics are averaged (instead of adding the relevant documents for all the topics and then computing the percentage value). Fig. 1(right) displays the average of the percentage values for all topics as function of the list depth for TREC 3, 7, 8, 12 (topics 1–50), 12 (topics 51–100) and 13, respectively. Fig. 1(right) shows that more than 50% of the total number of relevant documents are contained in the top 50 results; at least 80% of all relevant documents have been found for a list depth of 350 documents for all TREC years studied. This result should not come as a surprise since retrieval methods aim to place the relevant documents toward the top of their result lists.

The question arises whether the relevant documents found in the top 50 result lists are mostly the same or different documents. To answer this question, the average number of relevant documents found by different numbers of systems is computed for all topics. Specifically, the average number of relevant documents only found by one system is computed, followed by the number of documents found by two systems and so on, ending with the relevant documents retrieved by all systems. Fig. 2 displays the percentage of relevant

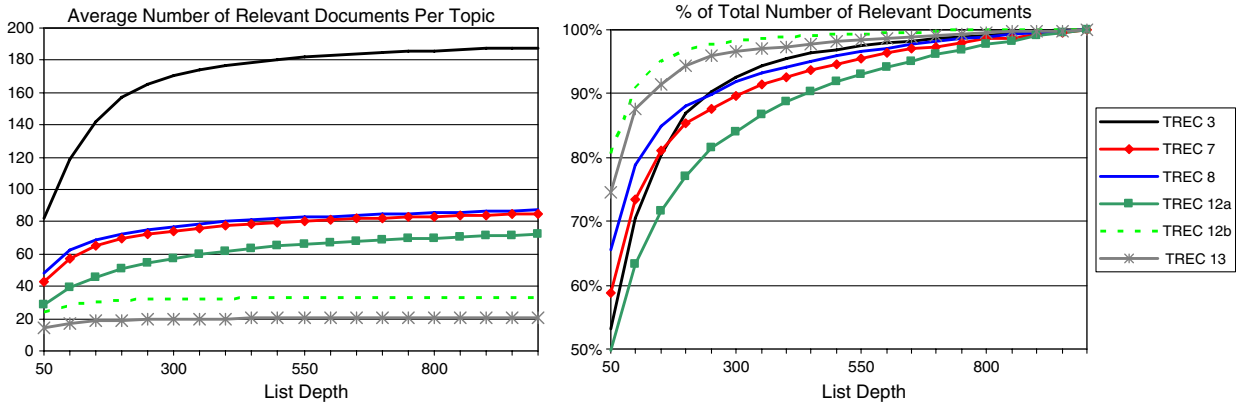

Fig. 1. Right: The average number of relevant documents found per topic in TREC 3, 7, 8, 12a (Robust: topics 1–50), 12b (Robust: topics 51–100) and 13 (Web), respectively, as a function of the list depth; left: the percentage of the total number of relevant documents found in TREC 3, 7, 8, 12a, 12b and 13, respectively, as a function of the list depth.

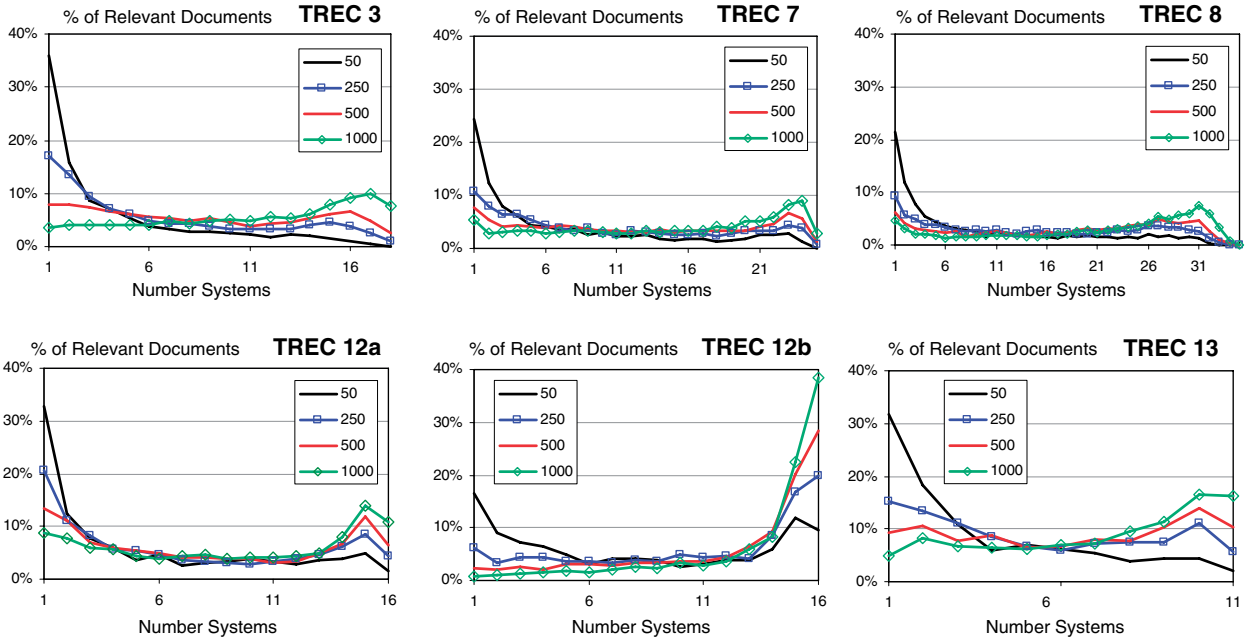

Fig. 2. The percentage of relevant documents found by a specific number of systems at the same list depth for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, respectively, if the top 50, 250, 500 or 1000 documents, respectively, are compared.

documents found by a specific number of systems at the same list depth for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, respectively, if the top 50, 250, 500 or 1000 documents, respectively, are compared. For a list depth equal to 50 (only the top 50 documents are compared), more than 30% of the relevant documents found at that level are only found by a single TREC 3, 12a or 13 system, respectively; for TREC 7, 8 and 12b, less than 25%, more than 20% and less than 20%, respectively, are found by a single system. Further, more than 50% of the relevant top 50 documents are found by five or less systems for TREC 3, 7, 8, 12a and 13, respectively; for TREC12b half of the relevant top 50 documents are found by seven or less systems. Clearly, predominately different relevant documents are retrieved by the different systems if only their top 50 documents are compared. As the list depth increases, Fig. 2 illustrates that an increasing percentage of the relevant documents is found by an increasing number of systems and the number found only by a single system decreases rapidly. For all TREC years studied, except TREC 12b, the highest percentage of relevant documents is found by almost all systems, only to decline because the worst performing systems retrieve very few relevant documents. If all the 1000 documents are compared, then 17 of the 18 systems being compared in TREC 3 find 10% of all the relevant documents; 24 of the 25 systems compared in TREC 7 find 9% and 31 out of 35 TREC 8 systems find more than 7% of all the relevant documents. Further, 15 of the 16 TREC 12a systems find 14% and 10 of the 11 TREC 13 systems retrieve 16% of the total number of relevant documents. As the list depth increases, the TREC 12b graph shows a similar pattern as the other TREC years studied, where the relevant documents are progressively found by more systems, but a greater percentage of relevant documents is found by all systems (i.e., almost 40% of the total number of relevant documents if all the 1000 documents are compared). Finally, for all TREC years studied, except TREC 12a, 5% or less of all the relevant documents are found by only a single system. In summary, Fig. 2 shows that the retrieval systems find similar relevant documents, but they do not find them in the same rank positions or at the same list depth levels. Thus, data fusion methods are well advised to consider all of the documents returned by the retrieval systems.

4.2. Authority Effect

In order to test the validity of the Authority Effect as the list depth is varied, it is necessary to compute the average number of both relevant and non-relevant documents that are found by different numbers of systems

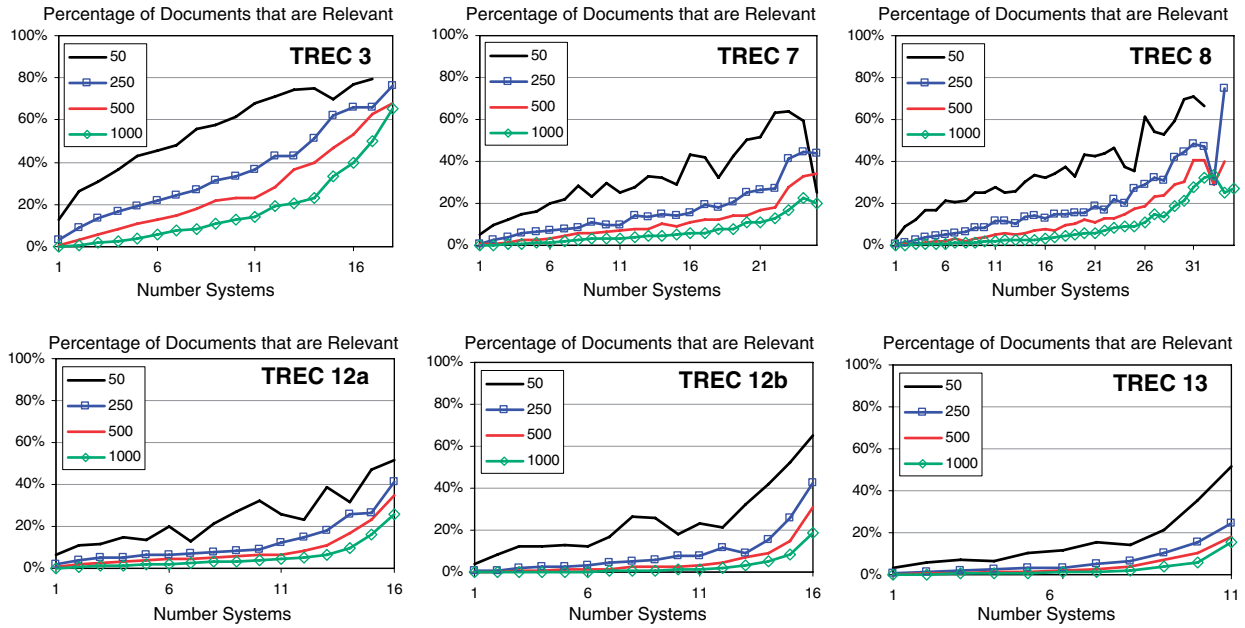

Fig. 3. The percentage of documents that are relevant as a function of the specific number of systems that found them at the same list depth for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, respectively, if the top 50, 250, 500 or 1000 documents, respectively, are compared.

at different list depth levels. As in the previous subsection, 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.

Fig. 3 displays the percentage of documents that are relevant as a function of the specific number of systems that found them for TREC 3, 7, 8, 12 (topics 1–50), 12 (topics 51–100) and 13, respectively, if the top 50, 250, 500 or 1000 documents, respectively, are compared. Fig. 3 shows that the percentage of documents that are relevant increases as the number of systems retrieving them increases, regardless of the list depth level. Specifically, for the top 50 documents, the percentage of documents that are relevant tends to increase linearly as the number of systems retrieving them increases and it then gradually increases exponentially as the list depth approaches 1000 documents for TREC 3, 7, 8 and 12a, respectively, where each of those TREC years has more than 70 relevant documents per topic on average. For TREC 12b and 13, which have 33 and 21 relevant documents per topic, respectively, the percentage of documents that are relevant increases exponentially for all list depth levels. As mentioned, TREC 3 contains more than twice as many relevant documents as TREC 7 or 8. This results in a much less significant drop of a document's maximal probability of being relevant when it is found by (almost) all systems in TREC 3 than in TREC 7 or 8 as the list depth increases.

In summary, Fig. 3 provides direct support for the Authority Effect – the probability that a document is relevant increases as more systems find it. Initially, this probability increases linearly, provided the average number of relevant documents per topic is greater than 50, and then gradually exponentially as the list level increases.

4.3. Ranking Effect

The question needs to be addressed if and how the Ranking Effect is affected by varying the list depth. If a document is found by multiple systems, then it will have multiple rank positions, which need to be averaged. First, the rank position is normalized so that the top document has a value of 1 and the very bottom document has a value of 1 divided by *ListDepth*, which is equal to the maximal number of documents currently being compared. Specifically, a document with rank position R will have a normalized rank value equal to

$1 - ((R - 1)/\text{ListDepth})$. For example, if ListDepth is equal to 50 documents, then the document in the 11th rank position will have a normalized rank value equal to 0.80. Second, these normalized rank values are averaged. In order to verify the Ranking Effect, it is necessary to compute the percentage of documents that are relevant (also called precision) as a function of the number of systems that find them, their average rank positions and the list depth used. For each specific number of systems, the data is aggregated using a range of consecutive average normalized rank values. The range or bucket size is equal to the list depth divided by 20 so that if the ListDepth is equal to 500, then 25 consecutive rank values are aggregated. As mentioned, the data from all the topics is then averaged and it is required that at least three topics have documents that are found by a specific number of systems so that a few data points can not introduce spurious effects in the precision calculation. Fig. 4 displays the plots of the precision or average percentage of documents that are relevant and are found by $1, 2, \ldots$ up to 35 systems in TREC 8 for the list depths equal to the top 50, 100, 250, 500, 750 and 1000 documents, respectively. The average normalized rank value increases from left to right in Fig. 4 for each *segment* that represents the documents found by a specific number of systems. For example, using a ListLevel equal to 50 documents, a document, which is found by 20 systems and placed in the 11th rank position in all of the 20 lists containing it, will have a normalized rank value of 0.80 and will be located toward to the right end of the segment with label "20" in Fig. 4, since it will have a value of 20.80 on the x-axis of Fig. 4.

The ascending "saw tooth" patterns in Fig. 4 demonstrate that as both the number of systems retrieving the same document and the average of their normalized ranking positions increases (i.e., the document moves toward the top of the result lists), the probability of the document of being relevant tends to increase. As the list depth increases, the "saw tooth" pattern shifts to the right as the documents found by few systems become increasingly less likely to be relevant and more documents are found by (almost) all systems. However, the situation is more complicated if a list depth of less than 250 documents is used. For example, if only the top 50 documents are compared, then a document, which is found by two to five systems and has a low average rank position, does have a greater probability of being relevant than a document with a high average rank position, which is the reverse result than predicted by the Ranking Effect. As the number of systems that find the same document increases, a high average rank position progressively implies a greater probability of being relevant. As the list depth level is increased to the top 100 documents, the same pattern as for a list depth of 50 can be observed, but the maximal value in a segment is decreased and the expected Ranking Effect starts to occur for a lower number of systems than if only the top 50 documents are compared. Once a list depth of 250 or more documents is used, then the regular Ranking Effect occurs in all the segments, where the segment that contains the documents found only by a single system represents a special case, since the probability of such documents being relevant is practically zero.

TREC 3, 7, 12a, 12b and 13 produce similar plots of the average percentage of documents that are relevant as TREC 8, but the total number of relevant documents found in a TREC year plays an important role with respect to the specific number of systems that need to have found a document for the regular Ranking Effect to occur. If only the top 50 documents are compared, Fig. 5 displays the percentage of documents that are relevant as a function of the number of systems that have found them and their average rank positions for TREC 3, 7, 8, 12a, 12b and 13 respectively. For TREC 3, 7, 8 and 12a, which on average retrieve more than 50 relevant documents per topic, the *reverse* Ranking Effect is observed for documents found by a low number of systems. The only difference between the TREC years studied is the specific number of systems, called *N_RankingEffect*, for which the regular Ranking Effect starts to emerge. For TREC 3, the Ranking Effect begins to manifest itself if more than 10 systems find the same document, whereas for TREC 7, 8 and 12a this number is lower. As mentioned, TREC 3 has more than double the relevant documents than TREC 7, 8, or 12a. This suggests that the total number of relevant documents affects the value of N_RankingEffect. Finally, as the list depth level is increased, N_RankingEffect becomes smaller and approaches one. For TREC 12b and 13, which have less than 50 relevant documents per topic, N_RankingEffect is equal to one and the regular Ranking Effect occurs at all list depth levels. In summary, the presented results provide direct support for Authority Effect at all list depth levels, whereas the average number of relevant documents per topic plays an important role for the Ranking Effect. If a TREC year has on average more than 50 relevant documents per topic, then the Ranking Effect only begins to emerge as more systems have found the same document and/or the list depth increases. However, if the average number of relevant documents per topic is less than 50 documents, then the Ranking Effects occurs as expected at all list depth levels.

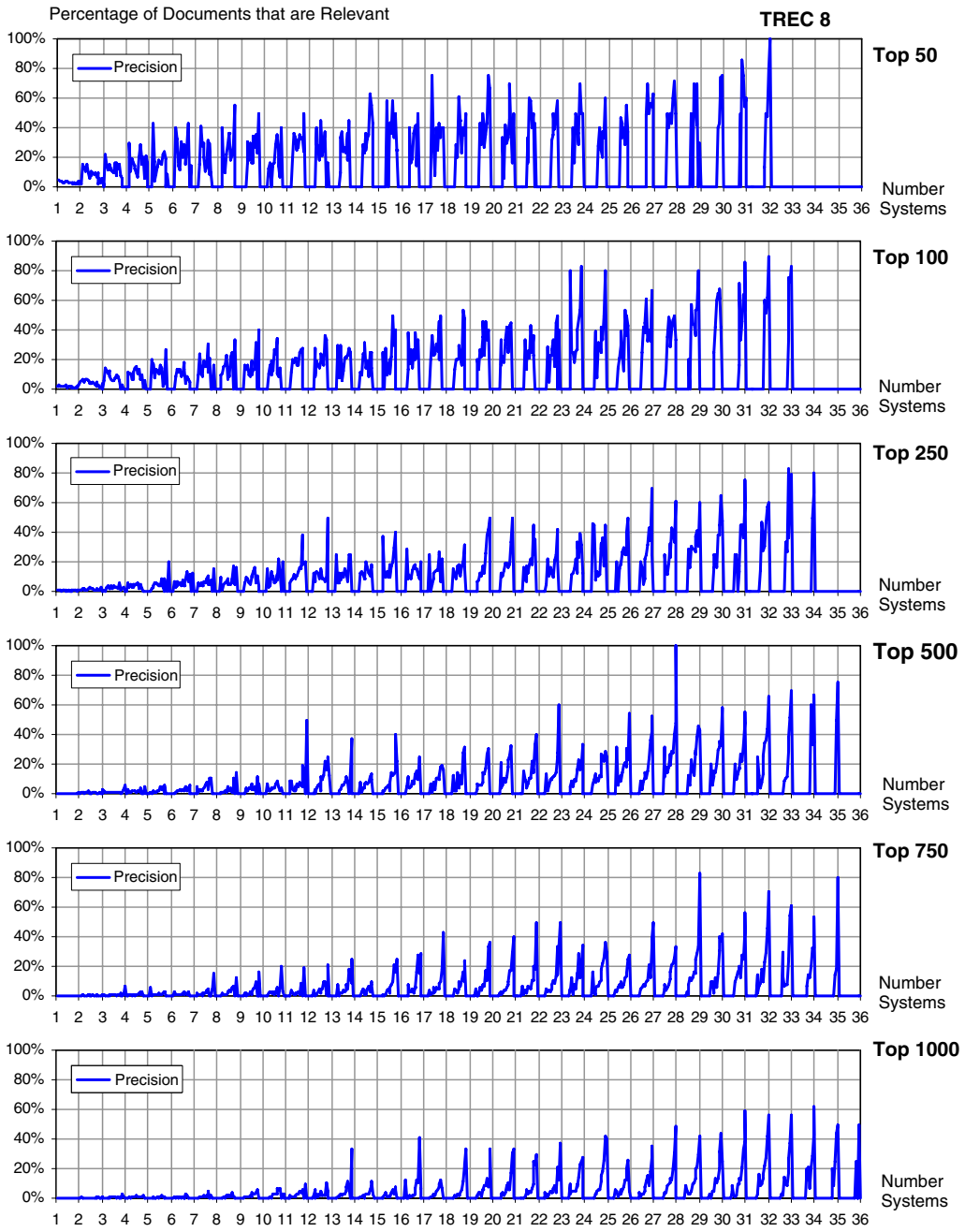

Fig. 4. Plots of the precision or percentage of documents that are relevant as a function of the number of systems retrieving them and the average of their ranking positions and the list depths of the top 50, 100, 250, 500, 750 and 1000 documents, respectively, for the 35 TREC 8 systems.

4.4. New documents found as list depth is increased

As mentioned, only the top 100 documents (or the top 125 documents for topics 51–100 in the TREC 12 Robust track) are pooled and evaluated for each topic. Since not all 1000 documents are examined, the question arises to what extend this could affect the Authority and Ranking Effects. Using TREC 3, 4 and 5 data, Zobel (1998) has investigated the number of unjudged relevant documents and estimated that for the topics

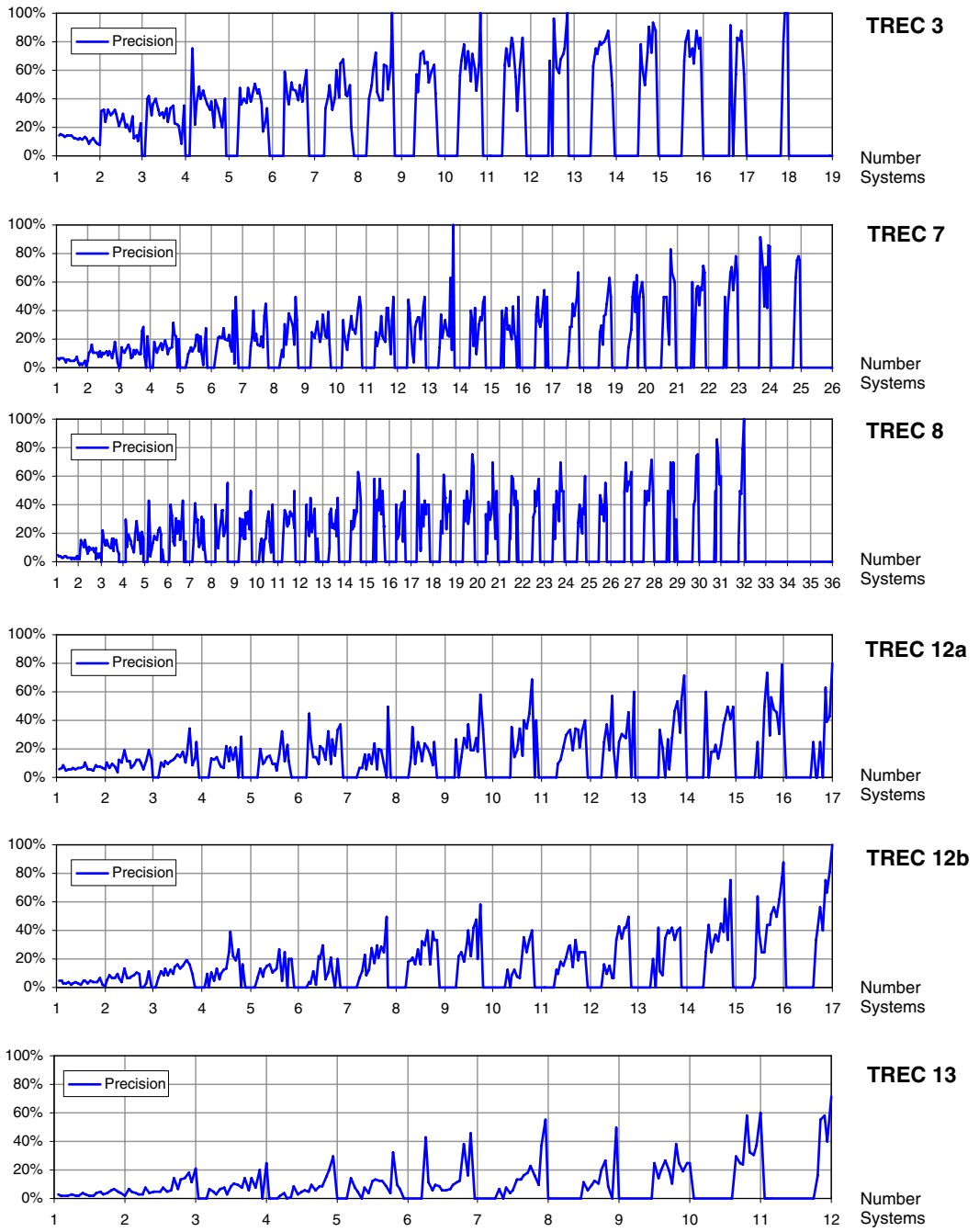

Fig. 5. The percentage of documents that are relevant as a function of the number of systems that have found them and their average rank positions for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, respectively, if only the top 50 documents are compared.

with very many relevant documents at best 50–70% of the relevant documents have been found when only the top 100 documents are examined. He arrived at this estimate by fitting a decreasing power function to the new relevant documents added for the list depths of 3–100 documents. Zobel concluded that this pooling bias does not significantly affect the accuracy of the performance evaluations of the retrieval systems participating in TREC. In this subsection, it will be shown that this pooling bias does not affect the validity of the Authority and Ranking Effects.

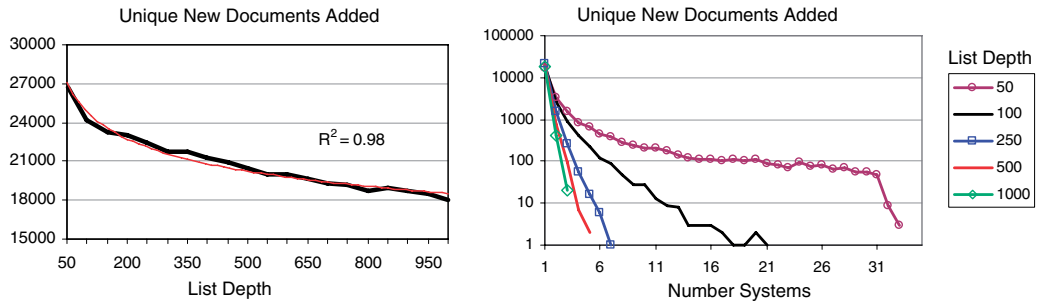

Fig. 6. Left: The number of unique new TREC 8 documents that are added if the list depth is progressively increased by 50 documents; right: the number of the unique new TREC 8 documents found by a specific number of systems for the list depths of 50, 100, 250, 500 and 1000 documents, respectively (using a log plot).

If the evaluators had examined more than the top 100 documents to identify relevant documents, then some of the documents with a rank position greater than 100 will not have been contained in any of the lists of the top 100 documents. How many such new documents are added can be computed as the list depth is increased. As shown in Fig. 6(left), the number of unique new documents added at each list depth level tends to decrease following a power law (as does Zobel's estimate of the new relevant documents added). Fig. 6(left) displays only TREC 8 data, but the graphs for the other TREC years analyzed in this paper exhibit similarly declining power functions for the new documents added at each list depth level. Next, it can be computed how many of these new documents are found by a specific number of systems as the list depth level increases. Fig. 6(right) displays how many of the unique new TREC 8 documents are found by a specific number of systems for the list depths of 50, 100, 250, 500 and 1000 documents, respectively. Fig. 6(right) shows that an increasing majority of these new documents are only found by a single system and very few new documents are found by more than three or four systems for list depths greater than 100 and approaching 1000 documents. The graphs for the TREC 3, 7, 12a, 12b and 13 years exhibit the very same pattern as shown for TREC 8 in Fig. 6. This implies that the unjudged relevant documents will be found by few systems at most and thus will not impact the results obtained for documents found by many systems. Further, as Zobel has estimated, the potential number of unjudged relevant documents decrease significantly as the list depth is increased and will therefore be "overshadowed" by the many non-relevant documents found by only one or few systems. Finally, if a relevant document is encountered as documents with rank positions greater than 100 are examined, then it is likely that it is contain in another system's list of the top 100 documents, since Fig. 2 shows that the relevant document contained in top 100 pool are found by an increasing number of systems.

In summary, the probability of encountering a relevant document decreases rapidly as document further down a result list are examined; if a relevant document is found, then it is likely that it is already contained in the top 100 pool. If the relevant document is not contained in the top 100 pool, then it will be only found by a single or very few systems and it will be "overshadowed" by the many non-relevant documents found by only a single or few systems. Hence, the impact of the undetected relevant documents with rank positions greater than 100 should not affect the Authority and Ranking Effects in a significant way.

4.5. Highly relevant documents

In TREC 12b (topics 51–100), the evaluators made three types of relevance judgments, "highly relevant", "relevant" and "not relevant", instead of the standard binary relevance judgments used in the other TREC years analyzed in this paper. This makes it possible to study if and how there are differences between the distributions of the "highly relevant" and "relevant" documents with respect to the Authority and Ranking Effects. Fig. 7(left) displays the percentage of "highly relevant" TREC 12b documents found by a specific number of systems at the same list depth if the top 50, 250, 500 or 1000 documents, respectively, are compared. Fig. 7(left) shows a similar pattern as the TREC 12b graphs in Fig. 2, but a greater percentage of the "highly relevant" documents is found by all systems and a lower percentage is only found by a single system. This

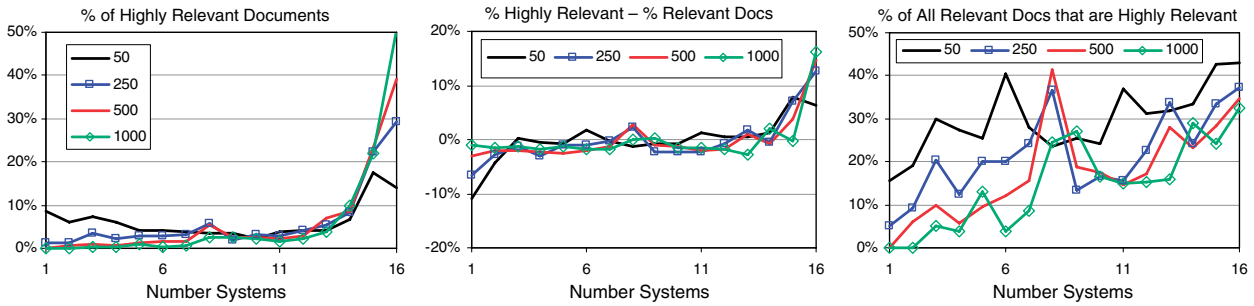

Fig. 7. Left: The percentage of "highly relevant" documents found by a specific number of systems at the same list depth for TREC 12b (topics 51–100) if the top 50, 250, 500 or 1000 documents, respectively, are compared; middle: the difference between the percentage of "highly relevant" and "relevant" TREC 12b documents found by a specific number of systems; right: the percentage of all relevant TREC 12b documents that are "highly relevant" and found by a specific number of systems at the list depth levels of the top 50, 250, 500 and 1000 documents, respectively.

difference is clearly illustrated in Fig. 7(middle), which displays the difference between the percentage of the "highly relevant" and "relevant" TREC 12b documents found by a specific number of systems. Finally, Fig. 7(right) shows the percentage of all relevant TREC 12b documents that are "highly relevant" and found by a specific number of systems at the list depth levels of the top 50, 250, 500 and 1000 documents, respectively. As the number of systems that find a relevant document increases, the percentage of documents that are "highly relevant" documents tends to increases as well, reaching a value greater than 30% for the relevant documents found by all systems irrespective of the list depth used.

5. Discussion and implications

As mentioned in Section 2, the major data fusion methods CombMNZ and CombSUM make use of the Ranking Effect, because they sum the normalized rank positions and 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 normalized rank positions are added. However, CombMNZ multiples CombSUM by the number of systems that find a document to make the Authority Effect dominant. Now, the data presented in this paper has implications for the effectiveness of CombSUM and CombMNZ if the list depth is varied and a large number of relevant documents are retrieved in a TREC year (i.e., the TREC 3, 7, 8 and 12a (topics 1–50) years analyzed in this paper). For example, if only the top 50 documents are compared, then more than 50% of the relevant documents are found by five or less systems for TREC 3, 7, 8, and 12a, respectively (see Fig. 2). However, CombSUM and CombMNZ will promote documents found by five or less systems that have a high average rank position, whereas Figs. 4 and 5 show that such documents tend to have a lower probability of being relevant than documents with a lower average rank position. Thus, CombSUM and CombMNZ promote the wrong documents found by five or less systems, which is the great majority of documents returned by the different systems. For documents found by more than five systems, these two major fusion methods progressively assign high scores to documents that are more likely to be relevant, but the number of relevant documents found by an increasing number of systems becomes increasingly small. In TREC years with a large number of relevant documents, the "tide begins to shift" in favor of the CombSUM and CombMNZ as the list depth level is increased, since progressively the probability distribution of a segment increases from left to right (i.e., the higher the average normalized rank position, the greater the probability the document is relevant) and its shape approaches that of an exponential function. On the one hand, if a higher percentage of the relevant documents are found by a greater number of systems, then CombSUM and especially CombMNZ will now promote progressively more documents that have a higher probability of being relevant. On the other hand, if an increasing number of relevant documents are considered, then the maximal value within a segment tends to decline as the list depth level is increased. In conclusion, as the list depth level increases, the relevant documents are progressively found by more of the systems and the probability distribution within a segment increasingly approaches an exponential function, where the maximal value within a

segment declines. Retrieval systems assign relevance scores to documents and use these scores to rank the documents to produce a ranked result list, where the relevance scores tend to decrease exponentially (Fox & Shaw, 1994). If CombMNZ and ComSUM only use the rank positions to the fuse the result lists, then they encode the probability distribution within a segment using a linear function, which is equivalent to assuming that very many relevant documents are found by the different systems being compared. If CombMNZ and CombSUM have access to and use the relevance scores assigned to the retrieved documents by the retrieval systems in the fusion calculation, then the probability distribution within a segment will in effect approach an exponential function, which may decay more rapidly than is warranted based on the data presented in Figs. 4 and 5.

Lee (1997) observed that different systems searching the same database tend to retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents. Fig. 2 shows that retrieval systems tend to retrieve similar sets of relevant documents if all the 1000 documents are considered. Fig. 6 indicates that most of the retrieved documents are only found by a single system or very few systems. Fig. 3 shows that a document, which has been found by a single system or few systems, has a probability of being relevant that is practically zero, if all the 1000 documents are considered, whereas a much higher percentage of documents found by multiple systems is relevant. This implies that retrieval systems tend to return dissimilar sets of non-relevant documents.

The systems participating in TREC may adopt the most successful solutions of previous TREC years. This could lead to a decreased variability between the systems, which could cause the result sets to overlap more greatly and in turn affect the Authority and Ranking Effects. If the differences between systems are reduced and become minimal, then this should greatly increase the overlap between the non-relevant documents, which in turn would cause the Authority and Ranking Effects to be greatly reduced. However, Fig. 3 shows that the Authority Effect has a similar structure for the different TREC years and the differences in its strength can be explained by the different total number of relevant documents found by the systems in a TREC year. Furthermore, the systems participating in TREC 12 Robust had the opportunity to train their methods using the topics 1–50. If this training had caused the systems to be become very similar, then the Authority and Ranking Effects for TREC 12a should be much weaker than the Authority and Ranking Effects for TREC 7 or TREC 8. Instead, the effects are comparable in strength considering that TREC 12a contains fewer relevant documents than TREC 7 or 8 (see Fig. 1(right)) and the TREC 12a consists of a subset of TREC 6–8 topics that proofed especially difficult for systems in those TREC years. In summary, if the different total number of relevant documents found in a TREC year are taken into account, then the similar structure and comparable strengths of the Authority and Ranking Effects for TREC 3, 7, 8, 12a, 12b and 13, and the TREC 12a data in particular, suggest that there is still a great deal of variability between the different systems participating in a TREC year and that the Authority and Ranking Effects represent robust phenomena.

As stated, the goal of the research presented in this paper is to investigate how much can be learned from analyzing the overlap between the result sets of different systems without the need for specialized knowledge, such as with particular algorithms were employed by the systems in a TREC track and year. Now, subsets of retrieval systems tend to use similar text retrieval and analysis methods or strategies (Craswell & Hawking, 2004; Voorhees & Harman, 1994, 1998, 1999; Voorhees, 2003, 2004). Future research will explore how the degree of similarity and differences between the retrieval algorithms used by the retrieval systems, which are being compared, may affect the Authority and Ranking Effects.

Data fusion methods usually consider all the runs submitted by systems participating in a TREC track and year. This makes is necessary to select many random subsets from this large set of runs to test a fusion method's effectiveness. It has been found that some subsets lead to better retrieval results than other subsets (Lee, 1997). Researchers have studied whether is possible to predict which subsets of runs will lead to retrieval results that are superior to the retrieval performance of the input runs (Lee, 1997; Vogt & Cottrell, 1998; Wu & McClean, 2006). In this paper, only one run of the runs submitted by the same systems is considered. Specifically, the "best" run with highest mean average precision is used in this study (but any single run submitted by a system can be selected). This greatly reduces the number of runs that need to be compared, making it not necessary to select random subsets of the runs, and more importantly, it greatly reduces the noise introduced if multiple runs by the same system are included in the analysis (Wu & Crestani, 2003). Thus, no special criteria are needed to select the systems used in this paper, expect that each system can contribute only one run to be included in the analysis.

6. Conclusions

This paper analyzed the overlap between the search results of retrieval systems that participated in the ad hoc track in TREC 3, 7 and 8, the robust track in TREC 12 and the web track (distillation task) in TREC 13 to investigate how the Authority and Ranking Effects are affected as the number of documents in the result lists, called the list depth, is varied. The selected TREC tracks and years made it possible to investigate both effects in diverse and difficult settings. First, it was shown that more than 50% of the total number of relevant documents are contained in the top 50 results and that at least of 80% of all relevant documents have been found for a list depth of 350 documents for all examined TREC years. Second, it was demonstrated that the retrieval systems find similar relevant documents, but they do not find them in the same rank positions or at the same list depth levels. In particular, it was shown that the relevant documents are progressively found by more systems and the number found only by a single system decreases rapidly as the list depth is increased. This suggests that data fusion methods are well advised to consider all of the documents returned by the retrieval systems. Third, it was shown that the Authority Effect is present at all list depths. Specifically, for the top 50 documents and the TREC 3, 7, 8 and 12a (topics 1–50) years that retrieve a large number of relevant documents, the percentage of documents that are relevant tends to increase linearly as the number of systems retrieving them increases and it then gradually increases exponentially as the list depth approaches 1000 documents. For TREC 12b (topics 51–100) and 13, which retrieve on average less than 50 relevant documents per topic, the percentage of documents that are relevant increases exponentially for all list depths. Further, TREC 3 has a much less significant drop in a document's maximal probability of being relevant when it is found by (almost) all systems than in TREC 7, 8, 12a, 12b or 13 as the list depth level increases, since TREC 3 contains at least more than twice as many relevant documents as the other TREC years. Fourth, it was shown that the Ranking Effect is present at all list depth levels, but if the systems in the same TREC year retrieve a large number of relevant documents, then the regular Ranking Effect only begins to emerge as more systems have found the same document and/or the list depth increases. Specifically, if a list depth of less than 250 documents is used, then the reverse Ranking Effect occurs, since a document, which is found by few systems and has a low average rank position, has a greater probability of being relevant than a document with a high average rank position. Fifth, it was demonstrated that the Authority and Ranking Effects are not an artifact of how the TREC test collections have been constructed (i.e., only top 100 documents are pooled and examined to identify relevant documents). Specifically, an increasing majority of new documents, which are found at a specific list depth level, are only found by a single system and very few new documents are found by more than three or four systems for list depths greater than 100 and approaching 1000 documents. This implies that the unjudged relevant documents will be found by few systems at most, only to be "overshadowed" by the many non-relevant documents found by few systems, and thus have no impact on the Authority and Ranking Effects for documents found by many systems. Sixth, it was shown that documents judged as "highly relevant" are found by more systems than documents judged as simply "relevant." It was also discussed how the effectiveness of the major data fusion methods CombSUM and CombMNZ is affected if the list depth is varied. If only a list depth of 50 or 100 documents is used and a large number of relevant documents are retrieved in a TREC year, then CombSUM and CombMNZ promote the wrong documents found by five or less systems, which is the great majority of documents returned by the different systems. As the list depth level is increased, the "tide begins to shift" in favor of the CombSUM and CombMNZ, since the Authority and Ranking Effects are now in full force. Finally, the question was addressed whether the Authority and Ranking Effects may be affected, since the systems participating in TREC may adopt the most successful solutions of previous TREC years. If the different total number of relevant documents found in a TREC year is taken into account, then the similar structure and comparable strengths of the Authority and Ranking Effects for TREC 3, 7, 8, 12a (topics 1–50), 12b (topics 51–100) and 13, and the TREC 12a data in particular, suggest that both effects represent robust phenomena.

Acknowledgements

The author would like to thank Nick Belkin and Paul Kantor as well as the reviewers of this paper for their helpful 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 http://trec.nist.gov/.

References

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 W. B. Croft (Ed.), *Advances in Information Retrieval* (pp. 127–150). Kluwer Academic Publishers.

Craswell, N., & Hawking, D. (2004). Overview of the TREC 2004 web track. *The Thirteenth Text Retrieval Conference (TREC-13), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

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. *2nd Annual Text Retrieval Conference (TREC-2), NIST, Gaithersburg, MD*.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *Proceedings of the 20th International 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. (2005). How the overlap between search results of different retrieval systems correlates with document relevance. In *Proceedings of the 68th Annual meeting of the American Society for Information Science and Technology (ASIST 2005)*.

Turtle, H., & Croft, B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems, 9*(3), 187–222.

Vogt, C., & Cottrell, G. (1998). Predicting the performance of linearly combined IR systems. In *Proceedings of the 21th International Conference on Research and Development in Information Retrieval (SIGIR'98)* (pp. 190–196).

Voorhees, E. (2003). Overview of the TREC 2003 robust retrieval track. *The Twelfth Text Retrieval Conference (TREC-12), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

Voorhees, E. (2004). Overview of TREC 2004 (TREC-13). *The Thirteenth Text Retrieval Conference (TREC-13), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

Voorhees, E., & Harman, D. (1994). Overview of the third text retrieval conference (TREC-3). *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

Voorhees, E., & Harman, D. (1998). Overview of the seventh text retrieval conference (TREC-7). *The Seventh Text Retrieval Conference (TREC-7), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

Voorhees, E., & Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). *The Eighth Text Retrieval Conference (TREC-8), Gaithersburg, MD, USA*. Washington: US Government Printing Office.

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).

Wu, S., & McClean, S. (2006). Performance prediction of data fusion for information retrieval. *Information Processing & Management, 42*(4), 899–915.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21th International Conference on Research and Development in Information Retrieval (SIGIR'98)* (pp. 307–314).