A Novel Class-Based Data Fusion Technique for Information Retrieval

Muath Alzghool
Al-Balqa' Applied University, Alsalt, Jordan
Email: alzghool@site.uottawa.ca

Diana Inkpen
University of Ottawa, Ottawa,Canada
Email: diana@site.uottawa.ca

Abstract— **Data fusion in information retrieval combines the results from multiple retrieval models or document representations. The achievement of data fusion technique is dependent on the quality of the inputs; classical data fusion techniques fail to improve the retrieval if the quality of the retrieval results varies from low to high quality. In order to tackle this problem, in this paper we address the issue of high variation among the retrieval strategies or document representations which affect the combination of their outputs. Our investigation on the MALACH speech collection – in which different segment representations are available – shows that neither the classical data fusion (CombSUM) nor the weighted version (WCombSum) improve the retrieval. We propose a novel class-based data fusion technique to deal with this issue. The segments retrieved by models based on different document representations are classified according to the quality of the segment into three classes: high, intermediate, and low quality class; then the similarity scores of each segment are fused using the classical CombSUM. Our experimental results show that the new technique is significantly better than CombSUM or WCombSUM in combing results with high quality variation.**

Index Terms—**Information storage and retrieval, searching spontaneous speech transcriptions, data fusion.**

I. INTRODUCTION

Conversational speech such as recordings of interviews or teleconferences is difficult to search through. The transcripts produced with Automatic Speech Recognition (ASR) systems tend to contain many recognition errors, leading to low Information Retrieval (IR) performance [1] unlike the retrieval from broadcast speech, where the lower word error rate did not harm the retrieval [2].

A large number of IR systems and retrieval strategies have been proposed and implemented in the last 30 years. All these approaches differ one from another in several issues such as the preprocessing process, the data representation, the weighting scheme and the similarity measure. There is a tremendous need to benefit from the strategies. One way to benefit from them is to combine their results by a data fusion technique.

Users tend to express their queries in various ways: sometimes they use more general terms, sometimes more specific terms, or a combination of both. IR systems need to be able to accommodate this variety of user needs; there is also variation among the collections (if it is a special collection like the one we use or a general collection like the news collection). Some retrieval models or weighting schemes perform better when the queries are general, others perform better when the queries are more specific, and others when a combination is available. In this paper we are looking for a system that will perform well in all these cases such as data fusion, where the system attempts to combine the results from multiple retrieval models.

Although the application of data fusion in IR has yielded good results in the majority of the cases, it has been noticed that its achievement is dependent on the quality of the input result lists [3-6]. Classical data fusion technique fail to improve the retrieval when the quality of the inputs varies from low to high quality; the presence of some poor-quality inputs (containing very few relevant documents in the top part of the list) causes a significant drop in the fusion performance

Lee [7] analyzed the overlap values of result sets from six different participants in TREC-3; he found that low overlap in non-relevant and high overlap in relevant documents is critical to improving effectiveness. We believe that the data fusion method should be able to combine the results that have high retrieval effectiveness with the results that have low retrieval effectiveness. Therefore, we propose a novel data fusion technique to fuse the results of different document representations, where the quality of the retrieval results varies from low quality to high quality.

We applied our data fusion techniques to Multilingual Access to Large spoken ArCHives collection (MALACH) [8] that used in the Cross-Language Speech Retrieval (CLSR) task at Cross-Language Evaluation Forum (CLEF) 2007. See Section 5 for a brief description of the collection.

The remainder of this paper is organized as follows. Section 2 is pointing to the most important work in model fusion. Section 3 describes the two IR systems that we used to provide candidate weighting schemes (retrieval

strategies) for our model fusion technique. Section 4 describes the data fusion technique proposed in this paper. Section 5 outlines the CLEF CL-SR test collection. Section 6 presents our experimental results. Finally, Section 9 presents conclusions and future work.

II. RELATED WORK

Model fusion combines the results from multiple retrieval models. Since different models may have different strengths, combining information extracted by multiple retrieval models can bring performance improvements. Fusion of retrieval results from different models for improving retrieval performance has been reported in works like [7, 9-12]. Retrieval results from different systems [10] or retrieval results using different document representations [11] were fused together for performance improvement.

Another way to differentiate data fusion methods is the way they compute the final score of documents. Some methods directly use the similarity values of the documents across the lists [9, 13], other consider their rank [3, 7], and others their probability of occurring in a predefined segment of the lists [14, 15]. In addition, some methods are based on the Social Choice Theory [16], he used pair wise contests of documents to determine their final score.

In general, a linear combination (CombSUM) of the retrieval results was found to be the simplest and most effective way for fusing multiple information sources in order to improve retrieval performance.

The application of data fusion in information retrieval has shown relevant results in the majority of the cases; nevertheless, it has been noticed that it is sensitive to several factors, such as the quality of the input lists[3-6].

III. SYSTEM DESCRIPTION

The weighting schemes for our fusion system were provided by two IR systems: SMART [17] and Terrier [18] .

SMART was originally developed at Cornell University in the 1960s. SMART is based on the vector space model of IR. We use the standard notation from SMART: the weighting scheme for the documents, followed by dot, followed by the weighting scheme for the query, where the schemes are abbreviated by the type of normalization (n means no normalization, c cosine, t idf, l log, etc.). We used the nnc.ntc, ntc.ntc, lnc.ntc, ntn.ntn, lnn.ntn, ltn.ntn, lsn.ntn weighting schemes[17]. We chose these schemes because they performed well on the training data in our last experiments[19].

Terrier was originally developed at the University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process.[18] We experimented with all the weighting schemes implemented in Terrier (BB2, BM25, DFR_BM25, DFRee, DLH13, DLH, IFB2, In_expB2, In_expC2, InL2, PL2, LemurTF_IDF, and TF_IDF).

IV. MODEL FUSION

A. CombSUM

Fox and Shaw [10] proposed several fusion methods for combining multiple scores. The most simple and effective one was called CombSUM, which sums up all the scores of a document, as in formula 1:

$$CombSUM = \sum_{i \in IR\ schemes} score_i \qquad (1)$$

where $score_i$ is the similarity score of the document to the query for the weighting scheme i which retrieved this document.

Since there are different weighting schemes from different systems, these schemes will generate different ranges of similarity scores, so it is necessary to normalize the similarity scores of the documents. Lee [7] proposed a normalization method by utilizing the maximum and minimum scores for each weighting scheme as defined by formula 2.

$$NormalizedScore = \frac{score - MinScore}{MaxScore - MinScore} \qquad (2)$$

B. Weighted CombSUM

When training data is available, many researchers experimented with updated versions of CombSUM, where a weight is assigned to each retrieval strategy according to performance on the training data. Then, they applied the determined fusion formula to the test data. This fusion method is called WCombSUM, represented by formula 3.

$$WCombSUM = \sum_{i \in IR\ schemes} W_{ik} * NormalizedScore_i \qquad (3)$$

where W_{ik} is a pre-calculated weight associated with each retrieval strategy, and the $NormalizedScore_i$ is calculated by formula 2 as described before.

In the literature, there are different ways to assign a weight (W_{ik}) for each retrieval strategy:

- Manually-weighted scheme [12, 20], where the researchers try different weight values for each retrieval strategy and select the best combination. We believe this technique is an unsystematic way to derive the weights.

- MAP-based weighted scheme [16, 21, 22], where the Mean Average Precision (MAP) score for each retrieval strategy on training data is considered as a weight for that strategy. This technique is simple and proves to be effective for some cases when there is no performance variation between the retrieval strategies on different data.

- MAP-Recall weighted scheme [19], where the MAP and the recall score are combined to derive the weight for each retrieval strategy so that the best weighting scheme contribute the most, and the others only support it.

In our experiments, we will use CombSUM and WCombSUM as baseline method, to compare it to our new technique. As a base case, we will consider the MAP

TABLE I. The retrieval results on MALACH collection using the weighting scheme DLH13 from the Terrier IR system on training data.

	Training	Test
Auto	0.1041	0.0735
Manual	0.3321	0.2560
Auto+Manual	0.2837	0.1606
CombSUM	0.2844	0.1953
WCombSUM	0.3272	0.2393

TABLE II. 11-level interpolated recall-precision values for the three experiments: Manual, Auto+Manual, and Auto. We show how to derive n and m, as explained in the text.

	Manual	Auto+Manual	Auto
Recall	Precision	Precision	Precision
0%	0.722	**0.697**	**0.424**
10%	0.577	0.504	0.247
20%	0.507	0.439	0.189
30%	0.435	0.353	0.146
40%	0.405	0.315	0.115
50%	0.353	0.282	0.091
60%	0.301	0.256	0.061
70%	0.242	0.200	0.041
80%	0.154	0.152	0.017
90%	0.090	0.088	0.023
100%	0.032	0.025	0.001

scores as the weights in the training phase for WCombSUM.

C. Class-Based Fusion

In this section we will discuss the case when we have different retrieval strategies and there are large differences in the effectiveness (significant difference), or we have one retrieval strategy and different representations for the documents, so that when we apply the retrieval strategy to the different representations, there are a significant differences among the different representations. Because of these differences, the basic fusion methods fail to improve the retrieval due to the noises from bad strategies or representations.

For example, the MALACH test collection contains 8104 segments from 272 interviews with Holocaust survivors and each segment contains different versions of automatic transcriptions, two sets of automatically-generated thesaurus terms, manually generated summaries, and manually-generated thesaurus terms. Each of them can be viewed as a representation for the segment. The first representation is when we index the automatically-generated data (Auto). The second one, is when we index the manually-generated data (Manual), and the third one is when we index the automatic and the manually-generated data together (Auto+Manual). If we apply any retrieval strategy to each representation, there are big differences among the representations, for example as shown in Table I, the MAP score for Auto, Manual, and Auto+Manual are 0.1041, 0.3321, and 0.2837, respectively. When the basic fusion methods like CombSUM or WCombSUM where the weights are the MAP scores on training data are applied, the MAP scores for the fusion methods are 0.2844 and 0.3272, respectively.

We are looking for a fusion technique that can handle the variations among the retrieval strategies or the document representations.

To achieve this goal, we will divide the retrieved documents from all the retrieval strategies or the document representations into three classes: the first one is expected to have the best precision values, the second one has intermediate precision values, and the last one has low precision; we will call these classes high, intermediate, and low class, respectively. Since the Manual experiment has the best MAP, we will assume the high class will have the top n documents from the

Manual experiment. The intermediate class will have the next m documents from Manual and the top m documents from Auto+Manual. Finally, the low class will have the remaining documents from Manual, Auto+Manual, and all the documents from Auto experiment. Note that the intersection between the three classes has to be mutually exclusive, i.e., if a document d appears in the top n documents from Manual and in the top m documents from Auto+Manual, d will be included in the high class, not in the intermediate class.

The next step shows how to estimate the values for n and m (n is the separation cut-off point between the high and the intermediate class, and m is the separation cut-off point between the intermediate and the low class). We use the evaluation of the three experiments on training data; for this stage we choose interpolated precision values at 11 recall points. To estimate n, for separating the high class from the intermediate class, we choose the maximum precision on Auto+Manual experiment, then find the level of recall that represents this value in the manual experiment, which is actually the same as looking at the length of the document list at the cut-off point; finally, we multiply this recall level by 1000 to calculate n (since the number of retrieved documents for each retrieval strategy is 1000, we take a portion of this number, which is proportional to the recall level). We use the same procedure for m; we chose the maximum precision on the Auto experiment, then find the level of recall on Auto+Manual and multiply it by 1000.

For example, Table II represents the precision at the 11-levels of recall for the three experiments mentioned in Table I. To estimate n, first we have to find the best precision in Auto+Manual, which is 0.697; then we have to find the level of recall that represents this value in the Manual experiment (0.1), and finally multiply this recall level by 1000; therefore, the estimated value for n is 100. We do the same thing for m; the maximum precision value in Auto is 0.424; the level of recall that represents this value according to the evaluation of the Auto+Manual experiment is 0.3; therefore, m is equal to 300. The high class will contain the top 100 documents

TABLE III. Some statistics about the number of terms and the number of tokens for the three experiments.

	Number of terms	Number of tokens	Average term frequency
Auto	13,605	1,711,684	125.8
Manual	7,131	278,717	39
Auto + Manual	15,884	1,990,401	125.3

TABLE IV. The average idf values, and number of missing search terms from title and description fields, for training (681 terms) and test (356 terms) topics

	IDF Training	IDF Test	Missing Training	Missing Test
Auto	1.22	1.08	28	8
Manual	1.75	1.74	27	9
Auto+Manual	1.22	1.05	10	5

from Manual; the intermediate class will contain the next 300 documents from Manual and the top 300 from Auto+Manual; finally, the low class will contain the remaining documents from Manual and Auto+Manual (600 and 700, respectively) and all the documents from Auto that were not included neither in the high class nor in the intermediate class. The three classes are mutually exclusive. In the above example, if one of the top 100 documents from Manual happens to be in the set of top 300 from Auto+Manual, then this document will be in the high class, not in the intermediate one.

The final step is to fuse the similarity scores of each document and to sort them in decreasing order in each class separately, then arrange the documents for the high class first, then the intermediate class, and finally the low class. To fuse the similarity scores, we could use CombSUM or WCombSUM. We have to normalize the similarity scores according to the maximum and minimum in each class. In our experiments, for any run that uses the class-based fusion, we will use the prefix "WC" before the method name, i.e., WCCombSUM.

V. THE CLEF CL-SR TEST COLLECTION

This section describes the data that we used. The MALACH collection contains 8104 "documents" which are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. Two ASR transcripts are available for this data, in this work we use the ASRTEXT2006B field provided by IBM research with a word error rate of 25%. Additional meta-data fields for each document include: two sets of 20 automatically assigned keywords determined using two different k-nearest neighbor classifiers (AK1 and AK2), a set of a varying number of manually-assigned keywords (MK), and a manual 3-sentence summary written by an expert in the field. A set of 63 training topics and 33 test topics were generated for this task. The topics provided with the collection were created in English from actual user requests. Topics were structured using the standard Text Retrieval conference (TREC) format of Title, Description and Narrative fields. For cross-language experiments, the topics were translated into Czech, German, French, and Spanish by native speakers. Relevance judgments were generated using search-guided procedure and standard pooling methods. See [8] for full details of the collection design.
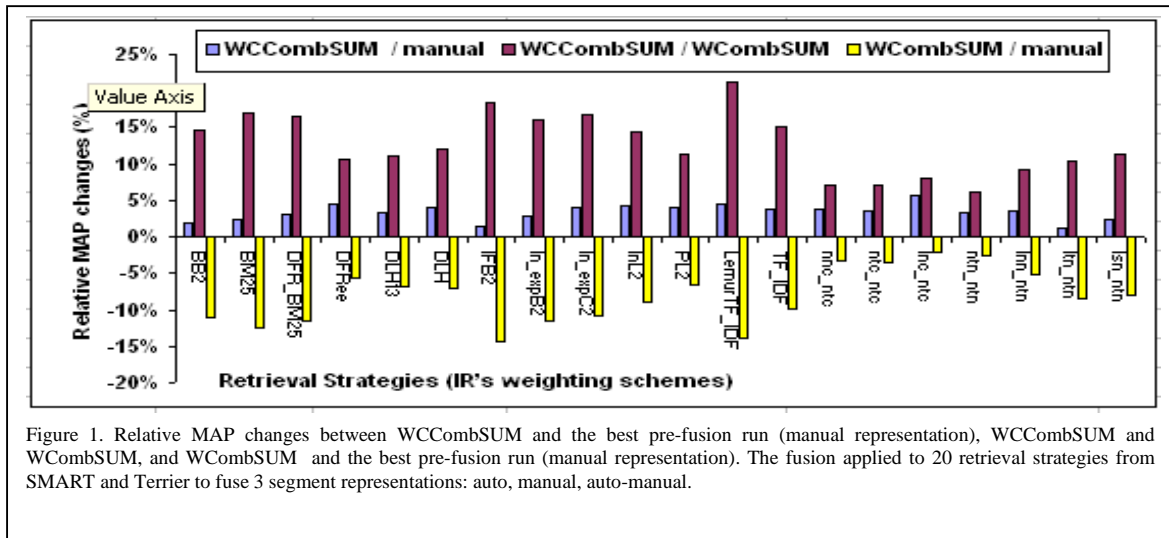
VI. EXPERIMENTAL RESULTS

The candidate retrieval strategies (weighting schemes) for our fusion system were provided by two IR systems: SMART [17, 23] and Terrier [18, 24].

We conducted three types of experiments, based on the fields which were indexed. In the first one, the automatic transcripts (ASRTEXT2006B), and two automatic keywords (AK1 and AK2) were used for indexing the documents; we call this experiment Auto. In the second experiment, we indexed the manual keywords and the manual summaries for each document; we named this experiment Manual. In the last experiment we indexed the automatic transcripts, the two automatic keywords fields, the manual summaries, and the manual keywords, we call this experiment Auto+Manual. The title and description fields from each topic are used as query. Table III shows some statistics about each experiment.

One interesting observation is that the number of terms (distinct words) in the manual fields is about half of the number of terms in the automatic fields. The number of tokens (total number of words) in the manual fields is about 16% of the number of tokens in the automatic fields. The average term frequencies are 39, 125, and 125 for Manual, Auto, and Auto+Manual, respectively. This ratio is very high: about four times more in the Auto fields. We also note that combining Auto and Manual brings about 14% of the terms to the Auto+Manual list of terms, which means that there is more information in the combined fields.

A. Manual Summaries and Keywords versus Automatic Transcripts

Experiments on manual keywords and manual summaries (Manual) available in the test collection showed high improvements over automatic transcripts and automatic keywords (Auto). The MAP score jumped from 0.0779 to 0.2727 on the test data. Also, if we indexed the Manual fields and the Automatic fields together (Auto+Manual), the MAP score jumped to 0.161, but it is far from the results on the Manual. This was also the case in the systems that participated in CLEF-CLSR. We are looking for a justification of why the difference is so big between the results of the Auto experiment and the Manual experiment, and why when we merge the Auto with Manual we do not reach the performance of the Manual fields. Since there are no manual transcripts available for the segments, we cannot

Figure 1. Relative MAP changes between WCCombSUM and the best pre-fusion run (manual representation), WCCombSUM and WCombSUM, and WCombSUM and the best pre-fusion run (manual representation). The fusion applied to 20 retrieval strategies from SMART and Terrier to fuse 3 segment representations: auto, manual, auto-manual.

know how the word error rate (WER) affects the retrieval.

We think that there are several factors that may affect the retrieval. The manual summaries are very concise representations of the segments; they tend to use different language than the segments. The automatic transcripts or the manual summaries cover the search terms from the training and test topics in different ways. Table IV counts the missing terms for each experiment in the training and test topics. We noticed that the number of the missing terms is approximately the same for Manual and Auto, and for Auto+Manual is approximately half the missing

number of terms from Manual or Auto. Therefore, we cannot consider the missing terms as the factor which affects the large difference in MAP score between Auto and Manual. Another factor could be related to the ability of the search terms to discriminate among the documents. The classic discrimination measure is the idf value for the search terms. Table IV shows the average idf for the training and test topics. We notice that the average idf for Auto and Auto+Manual is less than for Manual. Therefore, the topics ability to discriminate the documents in the Manual experiments is higher than for Auto or Auto+Manual. A last factor that we mention is

TABLE V. RESULTS (MAP SCORES, AND NUMBER OF RELEVANT DOCUMENTS RETRIEVED) FOR 20 WEIGHTING SCHEMES FROM SMART AND TERRIER, AND THE RESULTS OF THE FUSION METHODS (WCOMBSUM AND WCCOMBSUM), ON THE TEST DATA.

Weighting scheme	Auto		Manual		Auto+Manual		WCombSUM		WCCombSUM	
	MAP	Rel-Ret	MAP	Rel-Ret	MAP	Rel-Ret	MAP	Rel-Ret	MAP	Rel-Ret
BB2	0.0441	972	0.2699	1826	0.0970	1133	0.2402	1869	0.2752	1888
BM25	0.0567	1120	0.2490	1824	0.1404	1381	0.2182	1874	0.2551	1882
DFR_BM25	0.0580	1122	0.2558	1818	0.1408	1407	0.2261	1878	0.2635	1889
DFRee	0.0695	1298	0.2527	1822	0.1586	1697	0.2387	1897	0.2640	1900
DLH13	0.0735	1335	0.2560	1825	0.1606	1720	0.2384	1898	0.2647	1890
DLH	0.0719	1325	0.2460	1812	0.1606	1707	0.2287	1875	0.2560	1878
IFB2	0.0605	1080	0.2705	1824	0.135	1335	0.2320	1899	0.2747	1900
In_expB2	0.0657	1259	0.2727	1826	0.1537	1581	0.2416	1918	0.2805	1925
In_expC2	0.0700	1288	0.2704	1826	0.1551	1609	0.2409	1915	0.2812	1911
InL2	0.0629	1259	0.2575	1826	0.1521	1570	0.2346	1898	0.2685	1898
PL2	0.0730	1295	0.2510	1803	0.1575	1658	0.2347	1876	0.2613	1873
Lemur TF_IDF	0.0517	1146	0.2269	1814	0.1319	1425	0.1956	1867	0.2371	1874
TF_IDF	0.0651	1302	0.2525	1818	0.1452	1627	0.2277	1890	0.2620	1884
nnc_ntc	0.0779	1270	0.2190	1760	0.161	1698	0.2119	1837	0.2271	1845
ntc_ntc	0.0630	1235	0.2154	1776	0.1525	1623	0.2080	1845	0.2230	1873
lnc_ntc	0.0722	1269	0.2270	1784	0.1585	1667	0.2222	1865	0.2398	1879
ntn_ntn	0.0649	1250	0.2140	1792	0.1464	1643	0.2084	1857	0.2212	1867
lnn_ntn	0.0658	1284	0.2346	1789	0.1527	1684	0.2226	1880	0.2429	1897
ltn_ntn	0.0512	1166	0.2167	1785	0.1297	1511	0.1984	1844	0.2191	1880
lsn_ntn	0.0426	1028	0.1856	1787	0.1140	1376	0.1706	1795	0.1899	1832

the average term frequency, which is much larger in Auto and Auto+Manual (125) than in Manual (39), as previously shown in Table III.

Since the manual summaries and the automatic transcripts complement each other, each one brings new terms to the document structure as shown also in Table I. Mixing the two fields is supposed to improve the retrieval, in theory. From the results, it is clear that simple merging technique - during the indexing - does not help. A better way to combine or fuse the two fields during the indexing was addressed by [25].

In the next section, we will present the experimental results for the class-based fusion, which improve the retrieval, and benefits from the different information included in different segment representations.

B. Class-Based Fusion Experiments

We have applied our class-based fusion proposed in section IV.C to fuse the results from the three segments representations Auto, Manual, and Auto+Manual for each retrieval strategy (weighting scheme) from SMART or Terrier.

The baselines are the best retrieval run (the results from the Manual representation run) and the classical retrieval *WCombSUM*, where the weights in *WCombSUM* are represented by the MAP of each run on training data.

As shown in Table V and Figure 1 (see the previous page), the classical fusion technique (*WCombSUM*) does not improve the results comparing to the best run involved in the fusion process for the 20 retrieval strategies; but our method was better than the best run involved in the fusion for all the 20 retrieval strategy. For 15 out of 20 runs, the improvement was significant, based on a one-tailed Wilcoxon signed rank test with ($p < 0.05$). Also, our method was significantly better than the classical *WCombSUM* for all the 20 retrieval strategies. Based on number of relevant document retrieved, the *WCCombSUM* method is significantly better than the best run, *CombSUM*, and *WCombSUM*.

We conclude from our experiments that the information in meta-data like manual summaries and keywords complement the information contained automatic transcriptions and automatic keywords, and we could benefit from this feature to post-fuse the results of each representation and improve the retrieval.

VII. CONCLUSIONS

We have addressed the case when there are large differences in the effectiveness between the retrieval strategies or the document representations involved in the fusion, where classical techniques failed badly to improve the results. The solution was a class based method.

Finally, we have showed that meta-data complemented the error-full transcription, and we could benefit from the class-based fusion to improve the retrieval.

REFERENCES

[1] P. Pecina, P. Hoffmannová, G. J. F. Jones, Y. Zhang, and D. W. Oard, "Overview of the CLEF-2007 Cross Language Speech Retrieval Track," *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers,* vol. 5152, pp. 674-686, 2008.

[2] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *Proceedings of the RIAO: Content-Based Multimedia Information Access* PARIS, 2000.

[3] E. A. Fox and J. A. Shaw, "Combination of Multiple Searches," in *Proceedings of the Second Text REtrieval Conference (TREC 2)* Gaithersburg, Maryland, USA: National Institute of Standards and Technology (NIST), 1993.

[4] C. C. Vogt and G. W. Cottrell, "Predicting the performance of linearly combined IR systems," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* Melbourne, Australia: ACM, 1998.

[5] K. Ng and B. K. Paul, "Predicting the effectiveness of Naive data fusion on the basis of system characteristics." vol. 51: John Wiley & Sons, Inc., 2000, pp. 1177-1189.

[6] S. Wu and S. McClean, "Performance prediction of data fusion for information retrieval," *Information Processing & Management,* vol. 42, pp. 899-915 2006.

[7] J. H. Lee, "Combining multiple evidence from different properties of weighting schemes," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* Seattle, Washington, United States: ACM, 1995.

[8] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, B. Ramabhadran, M. Franz, and S. Gustman, "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* Sheffield, United Kingdom: ACM, 2004.

[9] B. T. Bartell, G. W. Cottrell, and R. K. Belew, "Automatic Combination of Multiple Ranked Retrieval Systems," in *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* Dublin,Ireland: ACM/Springer, 1994.

[10] J. A. Shaw and E. A. Fox, "Combination of Multiple Searches," *Proceedings of the third Text REtrieval Conference (TREC-3),* pp. 105-108, 1994.

[11] G. J. F. Jones, J. T. Foote, K. Spark Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* Zurich, Switzerland: ACM, 1996.

[12] K. Ng, "Information fusion for spoken document retrieval," in *Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference - Volume 04* Istanbul, Turkey: IEEE Computer Society, 2000.

[13] C. C. Vogt, G. W. Cottrell, R. K. Belew, and B. T. Bartell, "Using relevance to train a linear mixture of experts," in *The Fifth Text REtrieval Conference (TREC-6)* Gaithersburg, Maryland, 1997.

[14] G. Lebanon and J. Lafferty, "Cranking: Combining Rankings Using Conditional Probability Models On Permutations," in *Proceedings of the 19th International Conference on Machine Learning* Sydney, Australia, 2002.

[15] D. Lillis, F. Toolan, R. Collier, and J. Dunnion, "ProbFuse: a probabilistic approach to data fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* Seattle, Washington, USA: ACM, 2006.

[16] M. Montague, "Metasearch: Data Fusion for Document Retrieval," in *Computer science*. vol. Ph.D. Hanover: Dartmouth College, 2002.

[17] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management,* vol. 24, pp. 513-523, 1988.

[18] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier Information Retrieval Platform " *Lecture notes in computer science: Advances in Information Retrieval,* vol. 3408/2005, pp. 517-519, 2005.

[19] M. Alzghool and D. Inkpen, "Model Fusion Experiments for the CLSR Task at CLEF 2007," *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers,* vol. 5152, pp. 695-702, 2008.

[20] D. He and J.-W. Ahn, "Pitt at CLEF05: Data Fusion for Spoken Document Retrieval," *Accessing Multilingual Information Repositories,* vol. 4022/2006, pp. 773–782, 2006.

[21] D. He and D. Wu, "Toward a Robust Data Fusion for Document Retrieval," in *Proceedings of the 2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'08)* Beijing, China, 2008.

[22] M. Alzghool and D. Inkpen, "Cluster-based Model Fusion for Spontaneous Speech Retrieval," in *Proceedings of the Workshop on Searching Spontaneous Conversational Speech, SIGIR 2008* Singapore, 2008.

[23] C. Buckley, G. Salton, and J. Allan, "Automatic Retrieval With Locality Information Using SMART," *Proceedings of theText REtrieval Conferenc (TREC-1),* pp. 59-72, 1992.

[24] G. Amati and C. J. V. Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness* vol. 20. New York: ACM, 2002.

[25] G. J. F. Jones, K. Zhang, E. Newman, and A. M. Lam-Adesina, "Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech," in *SIGIR 2007 workshop: Search in Spontaneous Conversational Speech workshop* Amsterdam, 2007.