Data Fusion According to the Principle of Polyrepresentation

Birger Larsen and Peter Ingwersen
Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S, Denmark.
E-mail: {blar, pi}@db.dk

Berit Lund
Skills Creator, Vesterbrogade 149, 1620 København V, Denmark. E-mail: bl@skillscreator.com

We report data fusion experiments carried out on the four best-performing retrieval models from TREC 5. Three were conceptually/algorithmically very different from one another; one was algorithmically similar to one of the former. The objective of the test was to observe the performance of the 11 logical data fusion combinations compared to the performance of the four individual models and their intermediate fusions when following the principle of polyrepresentation. This principle is based on cognitive IR perspective (Ingwersen & Järvelin, 2005) and implies that each retrieval model is regarded as a representation of a unique interpretation of information retrieval (IR). It predicts that only fusions of very different, but equally good, IR models may outperform each constituent as well as their intermediate fusions. Two kinds of experiments were carried out. One tested restricted fusions, which entails that only the inner disjoint overlap documents between fused models are ranked. The second set of experiments was based on traditional data fusion methods. The experiments involved the 30 TREC 5 topics that contain more than 44 relevant documents. In all tests, the Borda and CombSUM scoring methods were used. Performance was measured by precision and recall, with document cutoff values (DCVs) at 100 and 15 documents, respectively. Results show that *restricted* fusions made of two, three, or four cognitively/algorithmically very different retrieval models perform significantly better than do the individual models at DCV100. At DCV15, however, the results of polyrepresentative fusion were less predictable. The *traditional* fusion method based on polyrepresentation principles demonstrates a clear picture of performance at both DCV levels and verifies the polyrepresentation predictions for data fusion in IR. Data fusion improves retrieval performance over their constituent IR models only if the models all are quite conceptually/algorithmically dissimilar *and* equally and well performing, in that order of importance.

Received May 16, 2008; revised December 1, 2008; accepted December 1, 2008

 DOI: 10.1002/asi.21028

Introduction

The principle of polyrepresentation has been developed as one of several consequences of a cognitive approach to Interactive Information Retrieval (IIR), as thoroughly discussed in Ingwersen (1996). From this perspective, any given retrieval model can be regarded as a representation of its designer(s)' retrieval ideas (his or her conceptual and algorithmic interpretation of IR) and is, in a cognitive sense, thus different from other retrieval models. According to the principle, different retrieval models retrieve different sets of information objects from the same collection of objects given the same information task, but some overlap of objects occurs (Croft & Thomson, 1987; Ingwersen & Järvelin, 2005; Larsen, Ingwersen, & Kekalainen, 2006; Wu & McLean, 2006). In polyrepresentation, the nature of their overlap depends on the conceptual/algorithmic interpretation of their "similarity." This is not the same as to state that a "relatively high" overlap of documents retrieved by symmetrically fused models *automatically* entails similarity of such models and a low retrieval performance, as studied by Ng and Kantor (2000). From a polyrepresentation perspective, a (relative) high overlap of documents may very well be an advantage. If the fused IR models are dissimilar (i.e., interpreting the information space from quite different perspectives), the overlap signifies high odds of relevant documents retrieved (Ingwersen & Järvelin, 2005, p. 208). The first attempt to utilize different IR models in combination for improved precision was by Croft and Thomson (1987) in their seminal I^3R retrieval system, fusing probabilistic and vector spaces models.

The principle of polyrepresentation operates with two types of (dis)similarity: "Cognitive dissimilarity" when fundamentally different IR models are in action and "functional difference" when the fused entities are based on different versions of the same fundamental retrieval model (Ingwersen, 1996). In this article, "conceptual/algorithmic (dis)similarity" refers to both types of polyrepresentation

and signifies the idea, assumptions, logic, and functionality behind a particular IR model. The (dis)similarity issue thus reflects the nature of automatic indexing rules, weights, retrieval strategies, modes of relevance feedback and query modification, and so on applied in IR models. The presented experiments test this principle and its predictive power by fusing three conceptually/algorithmically very different retrieval models with one model from the same algorithmic platform as one of the former.

Polyrepresentation principles have been shown to encompass different knowledge representations (interpretations) of information objects made by different actors such as authors or human indexers. Author interpretation is typically the full-text structures, image features, and references or outlinks (anchors). Metadata are added by human indexers or by indexing algorithms (interpretations). Studies have shown that when very different representations are combined and pointing to the same objects, the odds are indeed increased that these are relevant (Larsen, Ingwersen, & Kekäläinen, 2006; Skov, Larsen, & Ingwersen, 2006, 2008). The polyrepresentative combination improves retrieval performance over each of such representations in isolation.

Relevance feedback methods based explicitly on the principles of polyrepresentation also have been shown to improve retrieval performance (White, 2006). Finally, polyrepresentation has been proposed for describing a searcher's information-need situation in various forms (Ingwersen, 1996) and experimentally shown also to increase IR performance significantly by Kelly, Deepak, and Fu (2005) and Kelly and Fu (2007). Since the polyrepresentation principle is validated for documents, IR interaction, and searchers, we also wish to test its potentials on the remaining central component of IR: the retrieval platform.

According to Wu and McClean (2006), the ideal situation for data fusion is that all involved IR models (a) are equal in performance, (b) have very weak correlation with each other, and (c) of pairwise fusion have equal strength of correlation (p. 1964). Wu and McClean's "correlation" concept deals with the *relative size* of the overlap between the involved models, not its qualitative properties. The notion is thus associated, but not identical, to the polyrepresentation principle of cognitive (and functional) (dis)similarity between IR models. In our tests, we observe how combinations of equally performing, unequally performing, and cognitively dissimilar models perform compared to fusions of similar retrieval models at different document cutoff values (DCVs). Our tests serve to direct the fusion combinations according to a theoretical cognitive framework that may predict and explain why the fusions provide the observed outcomes.

The article is organized as follows. First, selected experiments of data fusion in IR that are central to the present context are discussed. The experimental setting of the tests is followed by the performance results with respect to the DCV at 100 and 15 documents, respectively. For each DCV, the restricted inner overlaps and traditional fusion experiments of polyrepresentative nature are analyzed. A discussion and suggestions for future work end the article.

Data Fusion in a Polyrepresentation Perspective Applied to IR Models

Data fusion has been investigated from several perspectives. Essentially, data fusion techniques attempt to produce a better performing end result than would each of the retrieval models in isolation. Basically, two different approaches to data fusion in IR have been tested. One approach is based on evidence combination; that is, combining the results based on the original retrieval scores of the fused retrieval models. Gaps in scores between the documents are thus taken into account during the fusion, but normalization of the different retrieval scoring systems must be carried out prior to combination. The other approach combines the rank position of the documents derived from the fused models. In this way, normalization is avoided.

Fox and Shaw (1994) were the first to test data fusion in IR; they introduced the CombSUM and CombMNZ multievidence document-scoring methods. They made use of the retrieval scores assigned to the documents in the original ranked output from each system taking part in the fusion. In CombSUM, the scores from each system are aggregated for the same retrieved document. In CombMNZ (Combining Multiplied Non-Zero scores), this sum is boosted by multiplying by the number of models that retrieve the document (i.e., the retrieval models that provide it with nonzero scores within a given range of documents). Linear combination methods also have been tested (Vogt & Cottrell, 1999). According to Beitzel et al. (2004) boosting, as in CombMNZ, might not improve retrieval effectiveness in the cases of fusing highly effective retrieval strategies (models):

> [The] reasoning for this lies in the fact that, because of the component strategies are known to be highly effective, it is fair to assume that the ranking they provide for their results is already of fairly high quality (i.e. relevant documents are likely to already be ranked higher than non-relevant documents). (p. 863)

In addition, the likelihood of merging in unique relevant documents, especially in high ranks, will tend to be quite small.

The perspective of normalizing the individual system scores prior to fusion was tried out by Lee (1997). This seminal paper investigated the characteristics of the fusion overlaps and tested the hypothesis that "Different [retrieval] runs might retrieve similar sets of relevant documents but retrieve different sets of non-relevant documents" (p. 268). In contrast to Fox and Shaw (1994), Lee found that the CombMNZ method outperformed the CombSUM fusion technique with a small margin. Lee made use of already performed TREC 3 retrieval runs from six IR systems in pairs. He found that his hypothesis was supported by the experimental results. The overlaps of relevant documents found in the pairwise combinations were twice the size of the overlaps of nonrelevant documents. Quite recently, Lillis, Toolan, Collier, and Dunnion (2006) tested random data fusion techniques in ways that can be translated to principles of polyrepresentation, also by means of TREC 5 data.

DOI: 10.1002/asi

The second alternative, to assign new artificial scores (e.g., based on the ranking information), was first carried out in the Borda fusion technique by Aslam and Montague (2001). This technique combines the ranking number of each document retrieved by each retrieval model. They also introduced the Condorcet-fuse technique, which associates to a voting procedure from social choice theory. Based on TREC 3, 5, and 9 retrieval runs from all systems combined, Aslam and Montague (2002) showed that the latter fusion algorithm outperformed the CombMNZ and Borda techniques; however, the Condorcet algorithm has not been tested further in connection with fusion in IR.

A third perspective is to attempt to predict the best performing fusion combinations prior to fusion, as shown by Vogt and Cottrell (1999) using linear regression. They made use of TREC 5 ad hoc retrieval result data in pairwise fusions of three retrieval models. In line with Lee's (1997) findings, Vogt and Cottrell concluded that not all fusions may outperform their individual components; only when almost equally well-performing retrieval models are fused may the fusion (in pairs) improve retrieval performance over the constituents. They regarded the TREC relevance pooling methods as a large-scale data fusion for IR. Ng and Kantor (2000) applied TREC 5 routing data, and Wu and McClean (2006) made predictions by means of overlap correlation measures. Ng and Kantor's experiments investigated the predictive power of output dissimilarity and a pairwise measure of similarity of performance in symmetrical data fusion (p. 1177). Wu and McClean's study demonstrated that versions of the same IR model have a very high overlap correlation and thus retrieve almost the same documents—a case to be avoided because nonrelevant documents then tend to be promoted in line with relevant ones. This view somehow contradicts the Lee's findings, but his tests involved different retrieval models. In summary, there seems to be a consensus that all the fused IR models should perform equally well. In pairwise data fusions, the CombMNZ work quite well and are commonly used, with the CombSUM of retrieval scores performing nearly as well. When random combinations are tested, results are mixed concerning which technique to use. Normalization of retrieval scores should be applied, and if not feasible, the Borda solution where rank scores are used seems to perform well.

Most data fusion experiments in IR, including the ones presented here, are based on retrieval results from already performed searches in TREC. The analysis or prediction methodologies in the present study are associated with the principle of polyrepresentation. The difference lies in the approach to prediction. While the former methodologies attempt to make predictions based on automated and quantitative perspectives of the overlaps, our approach is from a qualitative and conceptual/algorithmic perspective. In a polyrepresentation sense, the more evidence deriving from qualitatively different sources that point to a document, the higher the probability that that document is (highly) relevant. Thus, the relative size of overlap between fused IR models has a quite different meaning in polyrepresentation. A large overlap from conceptually/algorithmically very different IR models may be highly advantageous in a performance perspective—particularly with a greater number of retrieval models than in pairwise fusions. As in Lee's (1997) case, this is due to the issue of the locations of the (highly) relevant documents on the ranked lists of each fused IR model. In polyrepresentation, the assumption is that each model of very different conceptual/algorithmic nature pushes up relevant documents on their individual ranked output lists. Such documents are either retrieved in the fused overlap, and thus obtain high fusion scores by CombSUM, or they obtain substantial fusion scores by being uniquely located towards the top of the rank in the single models. Less or nonrelevant documents are assumed to be found lower on the single lists and hence receive less extensive fusion scores. This is the reason that we do not boost by means of the CombMNZ algorithm.

In data fusion based on polyrepresentation, there are fundamentally two ways of combining retrieval systems. One way is the *restricted overlap* fusion. Following this method, only the inner (disjoint) overlap of the fused models provides the result set to be ranked. Figure 1 illustrates this experimental situation with four different retrieval models—and their triple and pairwise disjoint overlaps as well as their inner "central cognitive overlap" formed by all four models (*Fuse4*). As the sets are disjoint, each of the 11 overlapping sets constitutes *separate* data fusion results, and the performance of each can be studied. Note that in the restricted fusion, each single original retrieval model only includes the documents that are "leftover;" that is, they do not include any of the documents in overlaps with other retrieval models. For instance, the performance of the original retrieval model COR is thus measured on the documents found in COR oval (1) *but not including* the already retrieved documents in *Fuse2* to *Fuse4* areas (white area, Figure 1). Their performance

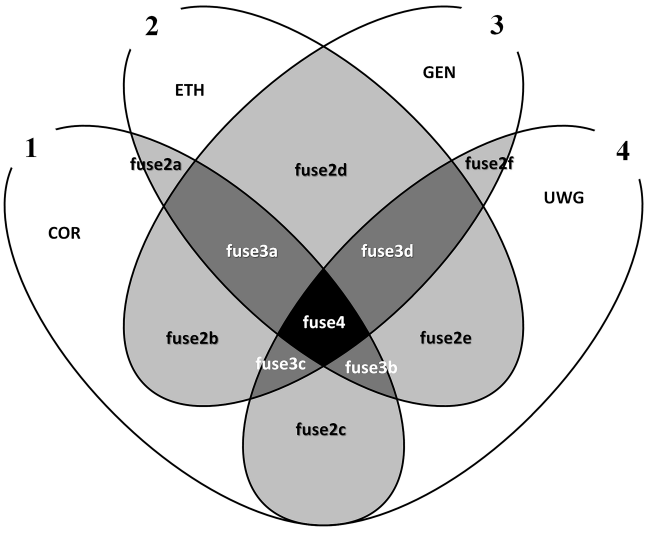

FIG. 1. Illustration of polyrepresentation of four different retrieval models' search results in the form of disjoint overlapping documents (variation of Ingwersen & Järvelin, 2005, p. 347; Lund, Schneider, & Ingwersen, 2006).

DOI: 10.1002/asi

TABLE 1. Fusion sample of TREC 5 Topic 254 over all 11 restricted fusion combinations at DCV100.

Combinations of TREC 5 IR models	Fusion Name	Retrieved documents	Relevant documents	Precision	Recall
ETH-GEN-UWG-COR	R-Fuse4	83	31	0.37	0.36
ETH-GEN-UWG	R-Fuse3d	85	32	0.38	0.38
COR-GEN-UWG	R-Fuse3c	90	33	0.37	0.39
COR-ETH-UWG	R-Fuse3b	84	31	0.37	0.36
COR-ETH-GEN	R-Fuse3a	100	32	0.32	0.38
ETH-UWG	R-Fuse2e	92	34	0.37	0.40
GEN-UWG	R-Fuse2f	100	32	0.32	0.38
COR-UWG	R-Fuse2c	92	33	0.36	0.39
ETH-GEN	R-Fuse2d	100	32	0.32	0.38
COR-ETH	R-Fuse2a	100	32	0.32	0.38
COR-GEN	R-Fuse2b	100	32	0.32	0.38

TABLE 2. Performance measures over four IR models and their combinations for 30 TREC 5 topics at DCV100. Italics signifying statistically significant results in relation to entities are marked by an asterisk, Friedman test ($\alpha = .05$). Best configurations are boldfaced.

	Restricted Fusion			Traditional Fusion		
Combinations of TREC 5 IR models	Fusion Name	Precision	Recall	Fusion Name	Precision	Recall
ETH-GEN-UWG-COR	R-Fuse4	***0.482***	0.295	Fuse4	0.448	0.262
ETH-GEN-UWG	R-Fuse3d	***0.472***	0.308	Fuse3d	***0.463***	**0.271**
COR-GEN-UWG	R-Fuse3c	***0.472***	0.301	Fuse3c	0.458	0.268
COR-ETH-UWG	R-Fuse3b	0.444	**0.311**	Fuse3b	0.437	0.256
COR-ETH-GEN	R-Fuse3a	0.392*	0.303	Fuse3a	0.401	0.235
ETH-UWG	R-Fuse2e	***0.457***	**0.341**	Fuse2e	**0.456**	**0.267**
GEN-UWG	R-Fuse2f	0.445	0.323	Fuse2f	0.425	0.249
COR-UWG	R-Fuse2c	0.425	0.317	Fuse2c	0.431	0.252
ETH-GEN	R-Fuse2d	0.414*	0.318	Fuse2d	0.411	0.241
COR-ETH	R-Fuse2a	0.391*	0.304	Fuse2a	0.395	0.231
COR-GEN	R-Fuse2b	0.385*	0.301	Fuse2b	0.381	0.223
UWG		**0.423***	**0.323**		**0.422**	**0.246**
ETH		0.404*	**0.323**		0.404	0.236
GEN		0.347*	0.279		0.353*	0.206
COR		0.343*	0.274		0.343*	0.201

is hence different from the performance obtained in unrestricted fusions (compare the restricted and traditional fusion columns in the last four rows in Table 2). Note also that in a restricted fusion of three systems, the inner overlap consists of a *Fuse3* version plus the document area labeled *Fuse4*.

In the second kind of data fusion based on polyrepresentation, disjoint sets are not studied separately. Individual retrieval models are instead combined as in *traditional data fusion*; that is, with underlying intermediate overlaps of documents filling up the ranked result list according to their fusion scores. A *Fuse3d* combination may hence, in principle, also contain documents from all the *Fuse2* to *Fuse4* areas and single systems, constituted by ETH, UWG, and GEN in Figure 1. In both experimental cases, one would intuitively choose fusion scoring methods such as Borda and fusion algorithms such as CombSUM, owing to their capacity for promoting the same document found by several (nonpairwise) fused IR models.

The present article discusses both types of polyrepresentative experiments with overlaps of retrieved documents from the 11 different logical data fusion combinations of four selected IR models from TREC 5 (see Tables 1 and 2).

Experimental Setting

The TREC 5 "ad hoc" track was selected as the test bed because of its variance of relevance intensity over its 50 ad hoc topics, Numbers 251 to 300. The number of relevant documents per topic ranged from 1 to 594. Six TREC 5 topics had more than 200 relevant documents while 12 topics had 100 to 199 relevant documents and 12 topics from 45 to 99 relevant documents. The remaining 20 topics contained 1 to 44 relevant documents. Sixty-one different retrieval models participated in TREC 5, and the mean average precision (MAP) over the 50 topics was on average .19, with a maximum of .317. A typical TREC 5 topic consists of a <title>, <narrative>, and a <description>, and all were rather short (i.e., 45 words on average for all three sections in total). In TREC 5, the document corpus consisted of approximately 4 GB text deriving from the *Wall Street Journal*, *AP Newswire*, *Federal Register*, *The Financial Times*, and

DOI: 10.1002/asi

the *Congressional Record* (TREC Disks 1–4). Participants could generate retrieval queries based on the <description> or on all three topic sections. Manually generated queries were allowed (Vorhees & Harman, 1997).

Based on the results from earlier experiments by Lund, Schneider, and Ingwersen (2006), the 30 most relevance intensive TREC 5 topics were selected for the present study. The minimum number of relevant documents in a topic was 45. Topics with few relevant documents would almost by definition perform badly according to polyrepresentation principles, as the chance of retrieving any relevant documents by several models would be very low compared to the risk of retrieving nonrelevant documents. In this respect, TREC 5 was ideal, with many topics having high relevance intensity. Unfortunately, from this perspective, from TREC 6 onward, topics with a large number of relevant documents were discharged from the experiments by the National Institute of Standards and Technology (Vorhees & Harman, 1998).

The experimental data derive from the results of the retrieval runs submitted for TREC 5 in 1996 and 1997 (i.e., the official TREC submissions, and not from reruns of retrieval models). Thus, the next stage was the selection of four TREC 5 retrieval models over the 30 topics with respect to high and equal performance as well as retrieval model (dis)similarity.

Just as it is important to include relevance intensive topics, it also is important to include high-performing IR models in the test; if low-performing models were included, there would be little chance of identifying relevant documents by fusion. Therefore, we chose the four best performing models over the 30 topics: Three of these were cognitively dissimilar models, and one was similar to one of the former. It would be ideal for the test if these models would have a very similar performance. This is a desirable property for the two cognitively similar models in particular. However, this is beyond our control as we rely on the official TREC submissions, and need to use high-performing models. One of the two best performing IR models was the UWG model from University of Waterloo, uniquely based on a new ranking principle (Clarke & Cormack, 1997). A second, but dissimilar, IR model was the ETH model from the Swiss Federal Institute of Technology, based on the classic SMART vector space platform (Ballerini et al., 1997). A third well-performing model, the Cornell University retrieval model COR (Buckley, Singhal, & Mitra, 1997), also was based on the classic SMART platform. The fourth model was a natural language processing based system (GEN) from a U.S. laboratory group (Strzalkowski et al., 1997). COR is hence similar to the ETH system and, like ETH, cognitively dissimilar to the other two IR models in the tests. The inclusion of COR and ETH thus allowed testing of the functional similarity issue against conceptually/algorithmically dissimilar IR models and made possible a falsification of the polyrepresentation hypothesis and predictions. From Tables 2 and 3, we can observe that ETH and UWG are the best performing of the four IR models.

In all tests, the top-1,000 results over the 30 TREC 5 topics from the four retrieval models were downloaded from the NIST Web site. The documents retrieved by each model were assigned an artificial ranking score (w), independent of the involved search models' own retrieval status values because it was not feasible to normalize the very different scales used. The scores assigned followed the Borda principle (Aslam & Montague, 2001) and ranged from 1,000 to 1 with respect to document positions 1 to 1,000 on the downloaded lists: $w = (1000 - R) + 1$, where R is rank number. In the case of document overlaps, the scores were summed as in CombSUM (Fox & Shaw, 1994). Thus, the higher the retrieval intensity (i.e., the number of IR models retrieving the same document), the higher the score for that document. Fusions and scores were calculated using a software tool developed for the purpose.

The fusions of the four models hold in their 11 logical combinations six in pairs (*Fuse2*), 4 triple combinations (*Fuse3*), and one full combination (*Fuse4*) (Figure 1). Tables 1 to 3, left-hand side, demonstrate the concordance between the fused IR model combinations (e.g., ETH-GEN) and the fusion name (e.g., *Fuse2d* or *R-Fuse2d*). In the article, the fusion name (*FuseX* or *R-FuseX*) is used.

In the first set of experiments on the *restricted fusion* (disjoint inner overlaps), combinations are named *R-Fuse* and measured at DCV = 100 by means of precision and recall, not by MAP, as less than 100 documents were often retrieved in the disjoint triple and quadruple restricted fusions (For double, triple, and quadruple fusions, the mean number of retrieved documents was 97, 86, and 74, respectively.) Table 1 provides an example of the number of retrieved and relevant documents and precision/recall calculations for Topic 254, which had a total of 85 relevant documents. For all 30 topics when less than 100 documents was retrieved, the calculations were based on the actual number of retrieved and relevant documents.

In the second set of polyrepresentation experiments applying *traditional fusion* methods, the combinations are named *FuseX* (e.g., *Fuse3d*). In Tables 2 and 3, the performance measures for both experiments made at DCV100 and DCV15 are precision and recall. Over all experiments, two *Fuse3* combinations hold three conceptually/algorithmically *very dissimilar models*, *Fuse3d* (UWG-ETH-GEN) and *Fuse3c* (UWG-COR-GEN). *Fuse3a* and *Fuse3b* each contain the two conceptually/algorithmically similar models (COR-ETH). Of the six *Fuse2* combinations, only one contains these two similar IR models: *Fuse2a*. Further, according to Table 2, the least performing retrieval model is COR. It forms part of *Fuse4*, $Fuse3a + b + c$ and $Fuse2a + b + c$.This distinction between (un)equality of IR performance and (dis)similarity of the fused models is important because according to Wu and McClean (2006), among others, less well performing retrieval models tend to diminish the performance result of a combination of well- and less well-performing models.

Results

Tables 2 and 3 show the results of the 2×2 sets of experiments for DCV100 and DCV15, with the restricted

DOI: 10.1002/asi

TABLE 3. Performance measures over four IR models and their combinations for 30 TREC 5 topics at DCV15. Italics signifying statistically significant results in relation to entities are marked by an asterisk, Friedman test ($\alpha = .05$). Best configurations are boldfaced.

	Restricted Fusion			Traditional Fusion		
Combinations of TREC 5 IR models	Fusion Name	Precision	Recall	Fusion Name	Precision	Recall
ETH-GEN-UWG-COR	R-Fuse4	*0.673*	**0.087**	Fuse4	*0.691*	*0.061*
ETH-GEN-UWG	R-Fuse3d	*0.671*	0.086	Fuse3d	***0.707***	***0.062***
COR-GEN-UWG	R-Fuse3c	*0.682*	0.087	Fuse3c	0.682	0.060
COR-ETH-UWG	R-Fuse3b	**0.687**	**0.089**	Fuse3b	0.678	0.059
COR-ETH-GEN	R-Fuse3a	0.673	0.085	Fuse3a	0.673	0.059
ETH-UWG	R-Fuse2e	***0.700***	**0.092**	Fuse2e	***0.696***	***0.061***
GEN-UWG	R-Fuse2f	0.671	0.083	Fuse2f	0.656	0.058
COR-UWG	R-Fuse2c	0.671	0.086	Fuse2c	0.667	0.059
ETH-GEN	R-Fuse2d	*0.693*	0.087	Fuse2d	0.689	0.060
COR-ETH	R-Fuse2a	0.649	0.082	Fuse2a	0.651	0.057
COR-GEN	R-Fuse2b	0.618	0.079	Fuse2b	0.616	0.054
UWG		0,600	0.079		0.604	0.053
ETH		**0.681**	**0,091**		**0,679**	**0,059**
GEN		0.564*	0.068		0.557*	0.049*
COR		0.614	0.072		0.606	0.053

fusion results to the left and the corresponding traditional fusion results to the right. In all experiments, precision and recall measures are used. The Friedman test (Siegel & Castellan, 1988) at $\alpha = .05$ was used to test for statistically significant differences.

Restricted Polyrepresentative Fusion at DCV100

The restricted combination *R-Fuse4* has a statistically significant advantage over all single IR models of which it is composed with respect to precision. Although disjoint inner overlap fusions rarely are experimentally tested, the *R-Fuse-4* result is in line with most traditional fusion experiments based on equally well-performing systems using the SUM scoring method (Fox & Shaw, 1994; Wu & McClean, 2006). A fusion of such systems commonly performs better than its constituents; however, our results also demonstrate that the intermediate *R-Fuse3c* + *d* combinations perform almost as well as *R-Fuse4*. Note that both these *R-Fuse3* combinations are made up of three algorithmically *very dissimilar retrieval models*. In contrast, *R-Fuse3a* + *b* combinations perform less well. They contain the two similar retrieval models (COR-ETH), with *R-Fuse3a* performing quite badly probably because this fusion also contains the less strong GEN model.

Table 2 further demonstrates that restricted fusions benefit from the strong combination of the two cognitively very dissimilar *and* well-performing models ETH and UWG, as demonstrated in the fusions *R-Fuse2e* and *R-Fuse3d*. The remaining five restricted double fusions all perform less well. *R-Fuse2c* + *d* + *f* combinations are performing better than are *R-Fuse2a* + *b* combinations probably owing to the fact that the former fusions all contain *dissimilar retrieval models* while the latter fusions suffer from (a) the inclusion of two algorithmically similar systems (*R-Fuse2a*: ETH-COR) or (b) a combination of the two least well-performing (although dissimilar) systems (*R-Fuse2b*: COR-GEN). An additional observation of interest is that *R-Fuse2b* performs better than its (dissimilar) constituents; in contrast, *R-Fuse2a*, including the least strong COR retrieval model, performs less well than one of its constituent models, ETH. This owes to both cognitive similarity *and* inequality in performance between the fused models. It is hence fair to state that dissimilarity of IR models is more influential in restricted fusions than is inequality of retrieval performance.

The reason for the observed difference in precision and recall of the single IR models between the restricted and the traditional fusion methods (Tables 2 and 3) is that in the restricted case, each model's precision/recall calculation is based on the restricted document area (white area, Figure 1). The disjoint overlapping documents (gray areas, Figure 1) are subtracted from the retrieved documents of each model prior to the calculation.

Traditional Fusion According to Polyrepresentation at DCV100

The unrestricted traditional fusions based on polyrepresentation follow a rather clear-cut pattern, but perform in general less well on precision than do the restricted fusions over the same combinations. In almost all fusion experiments carried out hitherto, fusions have been done pairwise or with many single systems, but rarely demonstrating the intermediate fusion results. Table 2 demonstrates such underlying combinations in fusion of four retrieval models as well as the top-level *Fuse4* result. Note the overall strength of the fusions of the conceptually/algorithmically *most dissimilar models* (*Fuse2e* and *Fuse3d*), with both performing better than *Fuse4* and with *Fuse3d* being statistically superior to the individual systems COR and GEN and indicatively better than UWG and ETH.

DOI: 10.1002/asi

Again, note the poor performance of *Fuse-3a* as well as *Fuse2a + b* compared to other intermediate fusions and the single IR models making up the combinations.

The overall observations do somewhat contrast claims in other fusion studies that fusions of many IR systems improve performance, as also observed by Ng and Kantor (2000) for pairwise fusions and Wu and McClean (2006) for multimodel fusions of TREC-5 data. We found that at DCV100, some particular *intermediate fusions* are superior in performance to their constituents as well as other fusion combinations. The best performing fusions are those made from conceptually/algorithmically very different and (almost) equally strong IR models. When more similar systems (and/or less well-performing IR models) are included in the fusions, they tend to perform less well, in general with the (dis)similarity causing most effect.

On the other hand, one can see that had the intermediate fusions *not* been carried out or shown, *Fuse4* alone performs as expected from earlier fusion experiments, by performing better than its constituent models (Table 2).

Restricted Polyrepresentative Fusions at DCV15

Table 3 demonstrates the pattern of restricted fusion according to polyrepresentation to the left. Different from Table 2, the pattern in Table 3, left-hand side, is fuzzier. Only *R-Fuse2e* of all the restricted fusions performs better than ETH as baseline on precision *and* recall, which is significantly the best single retrieval model at DCV15, followed by the other SMART-based system COR. All *R-Fuse3* and *R-Fuse4* combinations as well as the single IR models perform less well on precision than do *R-Fuse2d + e* and *R-Fuse3b* at this DCV level, regardless of their conceptual/algorithmic differences. At DCV15, GEN proves less strong than COR, hence *R-Fuse3b* outperforms the other *R-Fuse3* combinations even though the two similar systems COR and ETH both participate. The results of the restricted fusions show no significant variation across the 30 topics within the same combination.

Traditional Fusion According to Polyrepresentation at DCV15

In contrast to the restricted fusions, Table 3, the traditional fusions based on polyrepresentation demonstrate often higher precision, but lower recall scores and a steady trend identical to that at DCV100 for the same kind of fusion: The combinations of the cognitively *most dissimilar IR models* perform the best (*Fuse3d*: ETH-UWG-GEN; *Fuse2e*: ETH-UWG) with the four-model fusion, *Fuse4*, just behind. They clearly outperform all single IR models as baselines on precision and recall, of which again ETH performs the best, followed by COR and UWG at DCV15.

The same pattern displayed in the two tables for traditional fusion according to polyrepresentation is interesting since the conceptual/algorithmic *dissimilarity* of the retrieval models seems to play a more central role than their relative performance strength. The results indicate that the dissimilar models have been better able to push up the *same* relevant documents into top-15 (into a multi-overlap in *Fuse-3d*) compared to the performance-strong, but more similar models (e.g., *Fuse-3b*). However, very few results are statistically significant.

Discussion and Further Research

There are five important observations from the experiments. First, traditional data fusion based on principles of polyrepresentation and a cognitive framework for IR is verified and performs very well and significantly better than do single components as baselines at high as well as low DCVs. Hence, polyrepresentation can be applied as a data fusion principle, in line with traditional ad hoc fusion methods. The polyrepresentation assumption for fusion of IR models functions in line with what has been found to improve retrieval performance for polyrepresentation of documents (Skov et al., 2008), IR interaction (White, 2006), and searchers (Kelly et al., 2005).

Second, one may argue that the performance results of the fusions in comparison with the performance of the single constituent IR models are quite conservative. This is because the single models act as rather strong baselines since the top-100 documents from these models probably will have been completely assessed for relevance in the TREC pooling. However, in our fusion experiments, we use the top-1,000 results from existing TREC 5 runs, and there is a risk that some of the low-ranking documents were not part of the TREC pools and thus have not been assessed for relevance. They will be regarded as nonrelevant documents and be carried over into the fusions, although some might be relevant if assessed. With this issue of nonassessment in mind, Baillie, Azzopardi, & Ruthven (2008) questioned the appropriateness of reusing TREC runs for *new* experiments, owing to the increased uncertainty in system comparison when beyond the TREC pooling depth. Documents from the TREC test collections are not randomly assessed but derive from the list of top documents from the various participating IR models. A fair proportion of potentially relevant documents just below the top-ranked documents might hence never become assessed. Pooling the retrieved documents prior to assessment in TREC signifies a kind of polyrepresentation, but with the exclusion of the duplicates, which do not become weighted as in the present experiments based on CombSUM and Borda principles. In a sense, Table 1 illustrates how small the (restricted) overlaps actually can be between combinations of top-1,000 documents from just three to four IR models and the corresponding proportion of relevant documents. In our experiments, though, we rely on the existing runs, and the results are hence equally fair for all participating models at that time.

Third, the experiments demonstrate that restricted (disjoint overlap) fusions based on polyrepresentation do work as predicted by the principle at DCV100. Fusions of algorithmically dissimilar models outperform the single constituents

DOI: 10.1002/asi

and fusions made from similar systems. As expected, weak IR models tend to downgrade the fusion in which they participate. At a small DCV such as 15, the results are slightly fuzzier and thus less predictable with respect to impact of (dis)similarity versus performance (in)equality (e.g., *R-Fuse3b*). This is probably because the inner disjoint overlaps, Figure 1, are not sufficient alone to provide a steady and good retrieval performance. Some (highly) relevant documents are lurking outside those inner overlaps. Restricted fusion thus seems to be a special case of polyrepresentative data fusion that may provide higher *relevance density* in smaller sets of documents via higher precision *and* recall values (at longer DCV) than may traditional unrestricted polyrepresentative fusions, but in a less predictable way. If relevance feedback features were to be introduced, the missing relevant documents might be retrieved in ensuing runs, proving restricted fusion as a highly workable method for relevance density retrieval.

Fourth, traditional fusions of IR models based on polyrepresentation principles include indeed the underlying intermediate overlap documents according to their SUM scores. The (highly) relevant documents are pushed up towards the top rankings, and relevant documents lurking outside the inner overlaps, Figure 1, become included. The performance results become quite conclusive at both DCV levels, although not all are statistically significant: The best performing fusions are, at both DCV levels, those that combine the most conceptually/algorithmically dissimilar IR models, such as *Fuse2e* (ETH-UWG), *also* when the involved models demonstrate unequal retrieval performance, such as *Fuse3d* containing the least strong GEN IR model, plus ETH-UWG. The equality factor seems to be secondary in importance to dissimilarity.

Fifth, from the results it follows, as demonstrated in Tables 2 and 3, that a four or three IR model fusion may *not* necessarily perform better than a *Fuse2* combination—made across the *same* four IR models. In fact, had the tables not shown the intermediate fusion results, which are commonly never shown in fusion studies, one had simply observed the usual fusion result: that the *Fuse4* combinations constantly outperform their constituent IR models.

Which IR fusions score best among all possible combinations seems to depend on three factors in the following order: (a) the degree of conceptual/algorithmic dissimilarity between the constituent IR models, and (b) how equal and (c) well the component models perform. Typically, *Fuse4* (and *R-Fuse4* and some *Fuse3*) combinations have been shown to suffer simultaneously from all three factors in our experiments. Hence, we do not agree on the validity of the claim by Wu and McClean (2006) that a "high correlation" (i.e., a large overlap) between systems necessarily may lead to lower performance due to inclusion of too many nonrelevant documents via the overlap. The size of overlaps between several IR models (not just pairs) is always relative and, in itself, not the central issue; rather, the most important factor appears to be the *location* in the overlap, and in each fused system, of the relevant documents.

The strength of the principle of polyrepresentation lies in its predictive claim that other variables being equal, fusions of cognitively quite dissimilar IR models perform better than do fusions of related models or versions from the same algorithmic platform that are just functionally different. The conceptual/algorithmic differences between IR models imply quite different perspectives on the way of retrieving documents according to the model designers' ideas (e.g., the classification of retrieval models offered by Ingwersen & Järvelin, 2005, pp. 116–117). When such different perspectives meet in a polyrepresentation fusion and still point to the same documents (forming an overlap), those documents are probably highly relevant. The strength of data fusion in IR, from a cognitive theoretical perspective, is that relevant documents retrieved and ranked lower on the individual output lists of the fused models become part of the fusion and are pushed upward when found in the overlap. With very dissimilar (well-performing) IR models being fused, polyrepresentation predicts that the *same* nonrelevant documents are rarely retrieved equally high on the ranked lists. Thus, when found in an overlap, their combined fusion scores will be of lower value.

The experiments presented in this article did not falsify this claim but demonstrated its validity for unrestricted fusions. In future work, we will investigate how well the polyrepresentation principle and prediction work in data fusion in relation to graded relevance (Kekäläinen & Järvelin, 2002). The hypothesis is that data fusion based on polyrepresentation will retrieve a greater number of highly relevant documents within a given DCV compared to using equally well-performing, but not dissimilar, systems. It would be relevant also to test the application of different DCVs of the result lists used as input to the fusion (e.g., top-100 instead of top-1,000) to deal with the incomplete assessments introduced by the TREC pooling scheme.

A second set of experiments is intended to compare traditional IR methods to polyrepresentation principles applied simultaneously on document representations, fusion of IR models, and searcher statements of information requirements.

Acknowledgment

The authors thank Jaana Kekäläinen as well as one reviewer for their valuable comments, and the Nordic Research School in Library and Information Science (NORSLIS) for its travel and mobility support.

References

Aslam, J.A., & Montague, M. (2001). Models for metasearch. In D.H. Kraft, W.B. Croft, D.J. Harper, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001) (pp. 276–284). New York: ACM.

Aslam, J.A., & Montague, M. (2002). Condorcet fusion for improved retrieval. In C. Nicholas, D. Grossman, K. Kalpakis, S. Qureshi, H. van Dissel, & L. Seligman (Eds.), Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002) (pp. 538–548). New York: ACM.

DOI: 10.1002/asi

Baillie, M., Azzopardi, L., & Ruthven, I. (2008). Evaluating epistemic uncertainty under incomplete assessments. Information Processing & Management, 44(2), 811–837.

Ballerini, J.-P., Buchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schauble, P., & Wechsler, M. (1997). SPIDER retrieval system at TREC-5. In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 5th TREC Retrieval Conference (TREC-5) (pp. 217–228). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O., & Goharian, N. (2004). Fusion of effective retrieval strategies in the same information retrieval system. Journal of the American Society for Information Science and Technology, 55(10), 859–868.

Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC-5. In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 5th TREC Retrieval Conference (TREC-5) (pp. 105–118). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Clarke, C.L.A., & Cormack, G.V. (1997). Interactive substring retrieval (Multitext retrieval for TREC-5). In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 5th TREC Retrieval Conference (TREC-5) (pp. 267–278). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Croft, W.B., & Thomson, R.H. (1987). I^3R: A new approach to the design of document retrieval systems. Journal of the American Society for Information Science, 38(6), 389–404.

Fox, E., & Shaw, J.A. (1994). Combination of multiple searches. In D.K. Harman (Ed.), Proceedings of the 2nd TREC Retrieval Conference (TREC-2) (pp. 267–278). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec2/t2_proceedings.html

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, 52(1), 3–50.

Ingwersen, P., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), 1120–1129.

Kelly, D., Deepak, V., & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. In E.M. Vorhees & L.P. Buckland (Eds.), Proceedings of the 14th TREC Retrieval Conference (TREC 2005) (pp. 457–464). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec14/t14_proceedings.html

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. Information Processing & Management, 43(1), 30–46.

Larsen, B., Ingwersen, P., & Kekäläinen, J. (2006). The Polyrepresentation Continuum in IR. In Proceedings of the 1st International Conference on Information Interaction in Context (pp. 148–163). New York: ACM.

Lee, J.H. (1997). Analyses of multiple evidence combination. In N.J. Belkin, A.D. Narasimhalu, P. Willett, & W. Hersh (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997 (pp. 267–276). New York: ACM.

Lillis, D., Toolan, F., Collier, R., & Dunnion, J. (2006). ProbFuse: A probabilistic approach to data fusion. In E.N. Efthimiadis, S. Dumais, D. Hawking, & K. Järvelin (Eds.), Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006) (pp. 139–146). New York: ACM.

Lund, B., Schneider, J.W., & Ingwersen, P. (2006). Impact of relevance intensity in test topics on IR performance in polyrepresentative exploratory search systems. In R. White, G. Muresan, & G. Marchionini (Eds.), Proceedings of the ACM-SIGIR Workshop on Evaluating Exploratory Search Systems (EESS 2006) (pp. 42–47). Retrieved January 20, 2009, from http://research.microsoft.com/en-us/um/people/ryenw/eess/eess2006_proceedings.pdf

Ng, K.B., & Kantor, P. (2000). Predicting the effectiveness of naive data fusion on the basis of systems characteristics. Journal of the American Society for Information Science, 51(13), 1177–1189.

Siegel, S., & Castellan, N.J., Jr. (1988). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Skov, M. Larsen, B., & Ingwersen, P. (2006). Inter and intra-document contexts applied to in Polyrepresentation. In Proceedings of the 1st International Conference on Information Interaction in Context (pp. 763–770). New York: ACM.

Skov, M., Larsen, B., & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. Information Processing & Management, 44, 1673–1683.

Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Fang, L., Perez-Carballo, J., et al. (1997). Natural language information retrieval: TREC-5 report. In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 5th TREC Retrieval Conference (TREC-5) (pp. 291–314). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Vogt, C.C., & Cottrell, G. (1999). Fusion via a linear combination of scores. Information Retrieval, 1(3), 151–153.

Vorhees, E., & Harman, D. (1997). Overview of the 5th Text Retrieval Conference (TREC-5). In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 5th TREC Retrieval Conference (TREC-5) (pp. 1–20). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Voorhees, E., & Harman, D. (1998). Overview of the 6th Text Retrieval Conference (TREC-6). In E.M. Vorhees & D.K. Harman (Eds.), Proceedings of the 6th TREC Retrieval Conference (TREC-6) (pp. 1–24). Retrieved January 20, 2009, from http://trec.nist.gov/pubs/trec6/t6_proceedings.html

White, R. (2006). Using searcher simulations to redesign a polyrepresentative implicit feedback interface. Information Processing & Management, 42(5), 1185–1202.

Wu, S., & McClean, S. (2006). Improving high accuracy retrieval by eliminating the uneven correlation effect in data fusion. Journal of the American Society for Information Science and Technology, 57(14), 1962–1973.

DOI: 10.1002/asi