Ranking Fusion Methods Applied to On-line Handwriting Information Retrieval

Sebastiàn Peña Saldarriaga[1], Emmanuel Morin[1], and Christian Viard-Gaudin[2]

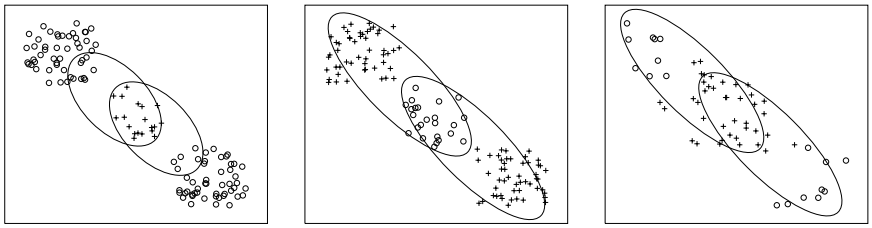
[1] LINA UMR CNRS 6241, Université de Nantes, France
[2] IRCCyN UMR CNRS 6597, École Polytechnique de l'Université de Nantes, France

Abstract. This paper presents an empirical study on the application of ranking fusion methods in the context of handwriting information retrieval. Several works in the electronic text-domain suggest that significant improvements in retrieval performance can be achieved by combining different approaches to IR. In the handwritten-domain, two quite different families of retrieval approaches are encountered. The first family is based on standard approaches carried out on texts obtained through handwriting recognition, therefore regarded as noisy texts, while the second one is recognition-free using word spotting algorithms. Given the large differences that exist between these two families of approaches (document and query representations, matching methods, etc.), we hypothesize that fusion methods applied to the handwritten-domain can also bring significant effectiveness improvements. Results show that for texts having a word error rate (WER) lower than 23%, the performances achieved with the combined system are close to the performances obtained with clean digital texts, i.e. without transcription errors. In addition, for poorly recognized texts (WER $>$ 52%), improvements can also be obtained with standard fusion methods. Furthermore, we present a detailed analysis of the fusion performances, and show that existing indicators of expected improvements are not accurate in our context.

1 Introduction

The use of ranking, or data fusion to combine document retrieval results has been addressed for several years by the Information Retrieval (IR) community [1–9]. The underlying assumption is that different retrieval techniques for the same query, retrieve different sets of documents [10]. By merging results from multiple systems, better retrieval effectiveness should be achieved for the same information need. However, as shown in Figure 1, there are several scenarios in which data fusion is ineffective or even more, harmful for retrieval performances. In particular, ranking fusion is not likely to improve results when the rankings involved are highly correlated [8].

Lee [11] claimed that different systems retrieve similar sets of relevant documents (+) but retrieve different sets of non-relevant documents (+), however when the overlap between relevant documents is too high, as in Figure 1(a), little or no improvement is to be expected [12]. In scenario 1(b), the common



(a) (b) (c)

Fig. 1. Several scenarios in data fusion for a given query. (a) high overlap between relevant documents (b) high overlap between non-relevant documents (c) ideal scenario.

non-relevant documents will dominate the fusion process and be promoted to higher positions, thus affecting the performances of the merged result. In the ideal scenario, shown in Figure 1(c)), the two sets share relevant documents but some documents appear only in one of them. In such a scenario, performance improvements are more likely to be observed when the non-common documents are merged into the fused set at a high rank [12].

To the best of our knowledge, the use of data fusion in handwriting retrieval has not yet been investigated. This study aims to investigate the effect of ranking fusion methods on handwriting retrieval. Current state-of-the-art in handwriting retrieval distinguishes two families of methods: recognition-based and recognition-free approaches. Since these methods use different document representations, query representations and retrieval algorithms, we hypothesize that the data fusion assumption holds in the handwritten domain. Thus, by combining the results of different handwriting retrieval methods, we can expect to improve retrieval effectiveness and leverage the strength of both method families.

The rest of this paper is organized as follows: Specificities and current methods in handwriting retrieval are reviewed in Section 2. Section 3 outlines the rank aggregation methods used in our experiments and the experimental methodology and data are described in Section 4. Experimental results are presented and discussed in Section 5, then conclusions are drawn in the final section.

2 Previous Works on Handwriting Retrieval

For several years, on-line handwriting was confined to the role of a convenient input method for PDAs, Tablet PCs, etc. With the recent evolutions of pen computers and digital pens, the production of on-line documents has become commonplace. These devices generate a series of two-dimensional coordinates corresponding to the writing trajectory as a function of time [13], called *on-line handwriting* or *digital ink*. As a result, algorithms for efficient storing and retrieval of on-line data are being increasingly demanded.

While the task of efficient retrieval of text documents has been addressed by researchers for many years, the retrieval of handwritten documents has been

addressed only recently [14–21]. Existing methods for handwritten document retrieval can be divided into recognition-based (noisy IR) [14, 18] and recognition-free (word spotting) approaches [15–17, 19–21].

Word spotting aims to detect words in a document by comparing a query word with the individual words in the document without explicit recognition, the query itself being either a handwritten text or an electronic string. The challenges with word spotting approaches is to deal with segmentation of handwritten texts into words and to cope with arbitrary writing styles.

In the case of noisy IR, a handwriting recognition engine is in charge of processing handwriting before the retrieval process. Then standard IR methods are applied to the output text. When carried out on the noisy texts obtained through handwriting recognition, IR techniques will be penalized by recognition errors. On the other hand, the critical point in word spotting is the proper selection of image features and similarity measures. Retrieval errors in word spotting are expected for words with similar shapes [16].

While being robust in determining which documents of a collection contain the keywords in the user query, word spotting robustness in satisfying information needs is not well established. In contrast, IR techniques carried out on transcribed texts should perform as well as standard methods as long as recognition is not very noisy [18]. Actually, we cannot tell *a priori* which method performs better than the others under all circumstances.

The interest of data fusion for handwriting retrieval is twofold. At first, an obvious reason is that fusion methods might improve retrieval effectiveness, since it is already the case in the text-domain. The second reason is that the combination involved two different levels of representation of handwritten texts: digital ink and transcribed texts.

According to the previous observations, we argue that by combining the results of different handwriting retrieval methods, we can improve retrieval effectiveness, while taking advantage of strengths of different techniques. A corpus of more than 2,000 on-line documents is used for experimental validation. Several fusion models were applied to rankings returned by different baseline retrieval algorithms, including both, recognition-free and recognition-based approaches. The fusion methods used are outlined in the section below.

3 Ranking Fusion Methods

Several ranking fusion methods have been proposed in the past in the IR literature [1–9]. Early works by Fox and Shaw [1] introduced a group of result merging operators such as CombMAX, CombSUM, and CombMNZ. Vogt and Cotrell [3] proposed a weighted linear combination method, in which training data is needed to determine the appropriate weight given to each input system. Wu & McLean [8] used correlation weights that do not require training data.

Data fusion can also be seen as a voting procedure where a consensus ranking can be found using Borda count [4], Condorcet method [6], Markov chains [7] and methods inspired from the social choice theory [9]. Manmatha et al. [5] proposed

to fit a mixture model consisting of an exponential and a Gaussian to the score distributions, then to average probabilities, thus minimizing the Bayes' error if the different systems are considered as independent classifiers; Bayesian fusion has also been explored [4]. Logistic regression has been employed successfully on one TREC collection [2].

As pointed out by Aslam and Montague [4], all of these methods can be characterized by the data they require: relevance scores or ordinal ranks, and whether they need training or not. In the present work we choose to use simple methods that require no training data, i.e. explicit user feedback. In particular the CombMNZ method has become a high-performance standard method in ranking fusion literature, whereas more complicated methods and weighting techniques exhibit mixed results.

In the following we will describe the methods used in our experiments. We adopt notational conventions from previous work [7]. These conventions are reviewed in Table 1.

Table 1. Notational conventions

Symbol	Definition
i	a document
τ	ranking of documents
$\tau\left(i\right)$	rank of document i
$\omega^{\tau}\left(i\right)$	normalized score of document i
$\mathcal{R}$	$\left\{\tau_1, \tau_2, \ldots \tau_{\left\|\mathcal{R}\right\|}\right\}$; set of rankings to fuse
$h(i, \mathcal{R})$	$\left\|\{\tau \in \mathcal{R} : i \in \tau\}\right\|$; number of rankings containing i
$s^{\hat{\tau}}\left(i\right)$	fused score of document i

The CombSUM operator corresponds to the sum of the normalized scores for i given by each input system, while CombMNZ is the same score multiplied by $h(i, R)$. Besides these two methods, a simple approach to combine estimated scores from each input system is to take the harmonic mean (CombHMEAN, Equation 1). Since the harmonic mean will tend towards the smallest score assigned to a document, any agreement between systems that is not supported by score agreement, i.e. similar scores are assigned to the same document, is thus minimized.

$$s^{\hat{\tau}}\left(i\right) = \frac{\left|\mathcal{R}\right|}{\sum_{\tau \in \mathcal{R}} \frac{1}{\omega^{\tau}(i)}} \tag{1}$$

Since the scores from the different systems are normalized $[0, 1]$, we propose to average the log odds, a method that has been shown to be effective in combining document filtering approaches [22] (CombODDS).

$$s^{\hat{\tau}}\left(i\right) = \frac{1}{\left|\mathcal{R}\right|} \sum_{\tau \in \mathcal{R}} \log \frac{\omega^{\tau}\left(i\right)}{1 - \omega^{\tau}\left(i\right)} \tag{2}$$

Since CombHMEAN and CombODDS are only defined for $\omega^\tau > 0$, i.e. they cannot deal with partial rankings, we assign an extremely small score to documents missing in one of the rankings to combine.

As suggested by Lee [11], using scores is equivalent to doing an independent weighting, i.e. without considering the whole result list. For this reason, the last two methods used in our experiments are based on ordinal ranks, i.e. considering the whole ranking. First we define a rank-derived score as follows:

$$r^\tau(i) = 1 - \frac{\tau(i) - 1}{|\tau|} \tag{3}$$

Then we use $r^\tau(i)$ with the CombSUM and CombMNZ operators. For convenience, the resulting methods will be called *RankCombSUM* and *RankCombMNZ*.

4 Methodology

The test collection used in our experiments is a handwritten subset of the Reuters-21578 corpus for text categorization (TC). An obvious difference between TC and IR test collections is that TC collections do not have a standard set of queries with their corresponding relevant judgments. However, we can use category codes to generate queries using relevance feedback techniques. Previous works reported IR experiments with the Reuters-21578 collection [23] using an approach similar to the one that is described here.

In the following, we consider the ground truth texts of our test collection randomly partitioned into two subsets of nearly-equal sizes:

- Q is the set used for query generation (1016 documents)
- T is the set used for retrieval (1013 documents)

Ranking Relevant Terms

In order to provide relevant terms for query generation, we used an adaptation of the basic formula for the binary independence retrieval model [24], which is the log odds ratio between the probability of term t occurring in a document labeled with category c ($p_{t,c}$); and the probability of term t occurring in a document not belonging to c ($q_{t,c}$).

$$score_{t,c} = \log \frac{p_{t,c} \times (1 - q_{t,c})}{(1 - p_{t,c}) \times q_{t,c}} \tag{4}$$

In practice, the probability estimates are $p_{t,c} = (x + 0.5)/(X + 1.0)$ and $q_{t,c} = (n - x + 0.5)/(N - X + 1.0)$, where a correction is applied to avoid zero denominators, and where x is the number of documents of c containing term t, X is the number of documents of c in Q, n is the number of documents containing t, and N is the number of documents in Q. Stopwords are not considered and words are stemmed. We keep the top 100 scoring terms as a basis for query generation as explained below.

Query Generation

The characteristic query q_c of a category c is generated using the Ide dec-hi [25] relevance feedback formula as follows:

$$q_c = \sum_{i=1}^{X} C_i - \sum_{j=1}^{X} S_j \tag{5}$$

Where C_i is the vector for the i-th document of c, S_j is the vector for the j-th document not belonging to c, and X is the number of documents of c. It is worth noting that the vector space dimension is 100, and that documents are indexed using unnormalized term frequencies. All the documents of c are used but only $|c|$ random negative samples. The queries generated are 5 terms long as suggested by Sanderson's results [23].

For each category in our corpus, the relevance feedback query is given in Table 2. By choosing a category, we can now perform a retrieval on T using the generated query. The documents tagged with the chosen category are considered as relevant.

Table 2. Generated queries for each of the 10 categories represented in the test collection. Query terms are stemmed.

Category	Query terms
Earnings	vs ct net shr loss
Acquisitions	acquir stake acquisit complet merger
Grain	tonn wheat grain corn agricultur
Foreign Exchange	stg monei dollar band bill
Crude	oil crude barrel post well
Interest	rate prime lend citibank percentag
Trade	surplu deficit narrow trade tariff
Shipping	port strike vessel hr worker
Sugar	sugar raw beet cargo kain
Coffee	coffe bag ico registr ibc

It is worth noting that for 6 categories, the name of the category is part of the query. Even though the generated queries are likely to be representative of their corresponding category, it is not clear if they make sense from a human perspective. Nevertheless, within the scope of our experiments, what is important is that all the retrieval methods perform the same task. Moreover, the query generation process can be seen as a single iteration of relevance feedback during an interactive retrieval session [23].

5 Results and Discussion

In this section, we present the experimental results for the ranking fusion methods described in Section 3 with the corpus and queries presented above.

5.1 Baseline Methods

Several existing retrieval methods were used as baseline systems to be combined. On the noisy IR side, three models, namely, **cosine**, **okapi** and **language modeling** (LM)[3] are used. Okapi parameters are set as they are usually set in the literature, and documents are weighted using the $tf \times idf$ measure in cosine retrieval.

On the word spotting side, we use **InkSearch®** (IS)[4], which is a stable and out-of-the-box system. It enables the searching of text in handwriting, and does not require training. Documents are scored by adding up the confidence scores of query word occurrences.

For the recognition-based methods, the recognition engine of MyScript Builder[5] is used. Recognition can be performed on a character-level basis (termed as *free*) or on a word-level basis (termed as *text*). The word recognition error rates for each of these two recognition strategies are reported in Table 3.

Table 3. Recognition error rates for *free* and *text* strategies on the handwritten dataset.

Recognition type	Word error rate
text	22.19%
free	52.47%

Unsurprisingly word-level recognition clearly outperforms character-level recognition [26]. Half the information is lost with the latter, while the former achieves low WER considering that no prior linguistic knowledge specific to these kind of documents is used. Results for baseline IR methods are reported in Table 4. In our work, we report the mean average precision (MAP) which provides a standard and stable evaluation metric [27].

The impact of recognition errors is obvious. In the case of documents recognized with the *text* strategy, a performance loss of roughly 5% is observed for every method, whereas this loss ranges from 15% to 20% with the *free* strategy.

5.2 Ranking Fusion Experiments

Table 5 shows the MAP obtained after fusion of the text-based approaches and IS. When using the *text* documents, significant improvements in performance are observed for *IS/Cosine* and *IS/Okapi* pairs with every fusion method. Combining IS and LM retrieval systematically degrades performance but never significantly.

[3] LM retrieval is based on the Kullback-Leibler divergence as implemented in the lemur toolkit, `www.lemurproject.org`

[4] InkSearch® is part of MyScript Builder SDK.

[5] MyScript Builder SDK can be found at `http://www.visionobjects.com/products/software-development-kits/myscript-builder`

Table 4. MAP obtained with the baseline methods. Columns indicate the type of recognition performed, the MAP obtained with the ground truth texts is also reported.

	Truth	Text	Free
Cosine	0.6887	0.6385	0.4980
Okapi	0.6989	0.6546	0.5005
LM	0.5589	0.4960	0.4101
IS	-	0.6547	0.6547

CombODDS and CombHMEAN perform as well as the standard methods. When the *text* documents are used, CombODDS and CombHMEAN achieves the best performance with *IS/Cosine* and *IS/Okapi* pairs respectively.

Rank-based methods are slightly outperformed by their score-based counterparts. Attempting to derive a cardinal score from an ordinal rank does not have a positive impact on the final results. Furthermore, this is highly questionable especially when the input rankings have different lengths [9], contain ties, and when the score distribution has a small statistical dispersion.

Table 5. Retrieval scores after fusion. Bold numbers indicate improvements with respect to IS and italic numbers indicate degradations in performances. An asterisk (*) indicates that the performance difference between IS and the combined method is statistically significant according to the Wilcoxon signed rank test at the 95% confidence interval.

(a) text

	Cosine	Okapi	LM
CombSUM	**0.6826***	**0.6933***	*0.6361*
CombMNZ	**0.6857***	**0.6933***	*0.6346*
CombODDS	**0.6871***	**0.6935***	*0.6346*
CombHMEAN	**0.6852***	**0.6940***	*0.6393*
RankCombSUM	**0.6775**	**0.6785**	*0.6351*
RankCombMNZ	**0.6808**	**0.6795**	*0.6347*

(b) free

	Cosine	Okapi	LM
CombSUM	**0.6782**	**0.6741**	*0.6451*
CombMNZ	**0.6760**	**0.6721**	*0.6428*
CombODDS	**0.6737**	**0.6719**	*0.6408*
CombHMEAN	**0.6710**	**0.6729**	*0.6410*
RankCombSUM	**0.6734**	**0.6692**	*0.6411*
RankCombMNZ	**0.6691**	**0.6644**	*0.6385*

Concerning the *free* documents, similar behaviour is observed. Once again, *IS/Cosine* and *IS/Okapi* pairs always produce improvements with every method, but these are not statistically significant. CombSUM achieves the best performance for every column in Table 5(b). It is worth noting that the performance difference between all the combining methods for *text* and *free* documents is only of 1% or 2%, while the difference between the individual methods ranges from 9% to 15%. Also, performances are slightly less degraded when the *free* documents are used in fusions with LM results.

5.3 Further Observations

In order to better understand why fusion techniques fail or succeed in bringing effectiveness improvements, we will try to examine different properties of the rankings. Lee [11] stated that improvements are expected when ranks have a greater overlap of relevant documents than of non-relevant documents. Later, Beitzel et al. [12] claimed that overlap rates were a poor indicator and proposed to relate fusion improvements to the Spearman's rank correlation coefficient.

Table 6 confirms that the unequal overlap property is not a good indicator of expected improvements. In the case of *text* documents, relevant/non-relevant overlap difference is about 64% for *IS/Cosine* and *IS/Okapi* pairs, and 56% for *IS/LM*. With *free* documents, a difference of 60% can be observed for *IS/Cosine* and *IS/Okapi* pairs, and 55% for *IS/LM*.

Table 6. Overlap and Spearman's correlation coefficient of recognition based systems with respect to IS.

(a) text

	IS/Cosine	IS/Okapi	IS/LM
R Overlap	95.71%	95.81%	78.20%
NR Overlap	31.73%	30.84%	22.04%
SR Correlation	0.7259	0.7366	0.7378

(b) free

	IS/Cosine	IS/Okapi	IS/LM
R Overlap	78.08%	76.90%	66.83%
NR Overlap	17.48%	16.85%	11.42%
SR Correlation	0.6439	0.6424	0.6753

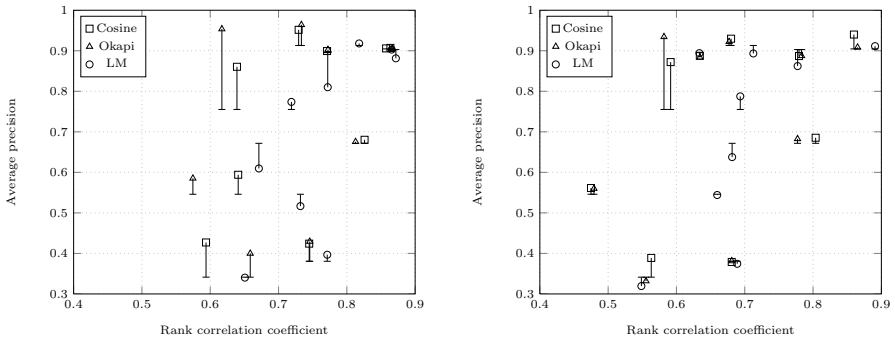
Next we examined the predictive quality of Spearman's rank correlation coefficient in our context. For each query, we computed the correlation coefficient between the rankings returned by IS and the text-based systems, and the average precision after fusion. The average correlation coefficient across queries is shown in Table 6.

Figure 2 shows the performance of CombMNZ as a function of the correlation coefficient. Disregarding the type of recognized documents, there is a positive Spearman correlation between all ranking pairs. Following Beitzel et al. [12], these correlations can be considered "moderate" to "strong" positive correlations, and any positive effects due to fusion should be minimized. However, we can observe in Figure 2 that for several queries, substantial improvement is achieved.

Furthermore, referring back to Tables 5 and 6, we can see that the best performances are obtained by combining ranks that exhibit higher correlation. Note that combining IS with LM retrieval always lead to performance degradation regardless of the correlation. Note also that the rank correlation coefficient does not take into account the relevance of documents in the rankings. This can explain why it cannot accurately predict expected improvements.

6 Conclusion

In this contribution, we have presented an empirical study on the combination of different approaches to on-line handwriting retrieval using data fusion methods.



(a) text (b) free

Fig. 2. Performance of fusion with CombMNZ as a function of Spearman's rank correlation coefficient. Each point represent a query, a point above the bar indicates an improvement with respect to IS, and a point below indicate degradations.

Experiments have been conducted with a handwritten subset of the Reuters-21578 corpus. Since this test collection does not contain a set of standard queries, we used relevance feedback techniques to generate queries, then we used category labels as relevance judgments. The corpus was divided into two subsets of equal sizes, the first one was used to generate the queries, and the second one for retrieval experiments.

The experimental results for individual baseline methods showed that recognition errors have an important impact on retrieval performances, losses ranging from 5 to 20% were observed when passing from clean digital texts to recognized documents. We also observed that score-based combining operators produces better retrieval performance than rank-based ones. The proposed methods, CombODD and CombHMEAN, perform as well as standard methods in the literature. Despite performance losses induced by recognition errors, data fusion improves retrieval performances in most configurations.

Further analysis of ranking fusion results showed that overlap rates [11] and Spearman's rank correlation coefficient [12] are not accurate indicators of expected performances after fusion. The baseline methods fused satisfy the unequal overlap property, yet degradations in performance can be observed. On the other hand, runs that have higher correlation coefficients achieve the best results. It has also been argued that ideally ranks that are being combined should be close in performance [8], however results obtained with the *free* documents suggest that this is not necessarily true.

The conclusions presented in this paper are not to be quoted out of the context of handwriting retrieval, and the systemic differences between the families of methods combined. We are aware that our experiments have several limitations, in particular those related to the size of the test collection, the number of queries, and their impact on the reliability of significance tests. The number of systems to combine can be seen as another limitation. To the best of our knowledge

there is no available test collection for IR in the on-line handwriting domain, however we think that further experimental validation using bigger databases and human-prepared queries and relevance judgments should be conducted.

Another direction that needs to be explored is the relationship between recognition errors, degradations in baseline scores and expected fusion improvements, models and measures are needed to describe more precisely their inextricable relationships. This will be the subject of future work.

Acknowledgments

This research was partially supported by the French National Research Agency grant ANR-06-TLOG-009.

References

1. Shaw, J.A., Fox, E.A.: Combination of Multiple Searches. In: TREC-2, proceedings of the 2nd Text REtrieval Conference. (1994) 243–252
2. Savoy, J., Le Calvé, A., Vrajitoru, D.: Report on the TREC-5 Experiment: Data Fusion and Collection Fusion. In: TREC-5, proceedings of the 5nd Text REtrieval Conference. (1997) 489–502
3. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. Information Retrieval **1**(3) (1999) 151–173
4. Aslam, J.A., Montague, M.: Models for Metasearch. In: SIGIR '01, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (2001) 276–284
5. Manmatha, R., Rath, T.M., Feng, F.: Modeling Score Distributions for Combining the Outputs of Search Engines. In: SIGIR '01, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (2001) 267–275
6. Montague, M., Aslam, J.A.: Condorcet Fusion for Improved Retrieval. In: CIKM '01, proceedings of the 11th International Conference on Information & Knowledge Management. (2002) 538–548
7. Renda, M.E., Straccia, U.: Web Metasearch: Rank vs. Score Based Rank Aggregation Methods. In: SAC 2003, proceedings of the 18th Annual ACM Symposium on Applied Computing. (2003) 841–846
8. Wu, S., McClean, S.: Data Fusion with Correlation Weights. In: ECIR 2005, Lecture Notes in Computer Science. Volume 3408. (2005) 275–286
9. Farah, M., Vanderpooten, D.: An Outranking Approach for Rank Aggregation in Information Retrieval. In: SIGIR '07, proceedings of the 30th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (2007) 591–598
10. Belkin, N.J., Cool, C., Croft, W.B., Callan, J.P.: The Effect of Multiple Query Representations on Information Retrieval System Performance. In: SIGIR '93, proceedings of the 16th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (1993) 339–346
11. Lee, J.H.: Analysis of Multiple Evidence Combination. In: SIGIR '97, proceedings of the 20th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (1997) 267–276

12. Beitzel, S.M., Jensen, E.C., Chowdury, A., Grossman, D., Frieder, O., Goharian, N.: On Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. Journal of the American Society of Information Science & Technology **50**(10) (2004) 859–868
13. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on Pattern Analysis & Machine Intelligence **22**(1) (2000) 63–84
14. Russell, G., Perrone, M., Chee, Y.M.: Handwritten document retrieval. In: Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition. (2002) 233–238
15. Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: CVPR 2003, proceedings of the IEEE Conference on Computer Vision & Pattern Recognition. (2003) 521–527
16. Jain, A.K., Namboodiri, A.M.: Indexing and retrieval of on-line handwritten documents. In: ICDAR 2003, proceedings of the 10th International Conference on Document Analysis & Recognition. (2003) 655–659
17. Rath, T.M., Manmatha, R., Lavrenko, V.: A search engine for historical manuscript images. In: SIGIR '04, proceedings of the 27th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (2004) 369–376
18. Vinciarelli, A.: Application of information retrieval techniques to single writer documents. Pattern Recognition Letters **26**(14) (2005) 2262–2271
19. Jawahar, C.V., Balasubramanian, A., Meshesha, M., Namboodiri, A.M.: Retrieval of online handwriting by synthesis and matching. Pattern Recognition **42**(7) (2009) 1445–1457
20. Terasawa, K., Tanaka, Y.: Slit style HOG feature for document image word spotting. In: ICDAR 2009, proceedings of 10th International Conference on Document Analysis & Recognition. (2009) 116–120
21. Cheng, C., Zhu, B., Chen, X., Nakagawa, M.: Improvements in keyword search japanese characters within handwritten digital ink. In: ICDAR 2009, proceedings of 10th International Conference on Document Analysis & Recognition. (2009) 863–866
22. Hull, D.A., Pedersen, J.O., Schütze, H.: Method Combination For Document Filtering. In: SIGIR '96, proceedings of the 19th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (1996) 279–287
23. Sanderson, M.: Word sense disambiguation and information retrieval. In: SIGIR '94, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (1994) 142–151
24. Robertson, S.E., Spärck Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science **27**(3) (1976) 129–146
25. Ide, E.: New Experiments in Relevance Feedback. In: The Smart Retrieval System. Prentice-Hall, Inc. (1971) 337–354
26. Perraud, F., Viard-Gaudin, C., Morin, E., Lallican, P.M.: Statistical language models for on-line handwriting recognition. IEICE Transactions on Information & Systems **E88-D**(8) (2005) 1807–1814
27. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: SIGIR '00, proceedings of the 23rd Annual ACM SIGIR Conference on Research & Development in Information Retrieval. (2000) 33–40