Information Fusion for Combining Visual and Textual Image Retrieval

Xin Zhou, Adrien Depeursinge
Geneva University Hospitals and University of Geneva, Switzerland
Medical Informatics Service, Geneva, Switzerland
xin.zhou,adrien.depeursinge@sim.hcuge.ch

Henning Müller
University of Applied Sciences Western Switzerland (HES–SO)
TechnoArk 3, 3960 Sierre, Switzerland
henning.mueller@sim.hcuge.ch

Abstract

In this paper, classical approaches such as the maximum combinations (combMAX), the sum combinations (combSUM) and the multiplication of the sum and the number of non–zero scores (combMNZ) were employed and the trade–off between two fusion effects (chorus and dark horse effects) was studied based on the sum of n maximums. Various normalization strategies were tried out. The fusion algorithms are evaluated using the best four visual and textual runs of the ImageCLEF medical image retrieval task 2008 and 2009. The results show that fused runs outperform the best original runs and multi–modality fusion statistically outperforms single modality fusion. The logarithmic rank penalization shows to be the most stable normalization. The dark horse effect is in competition with the chorus effect and each of them can produce best fusion performance depending on the nature of the input data.

1. Introduction

In the ImageCLEF image retrieval competition, multimodal image retrieval has been evaluated over the past seven years. For ICPR 2010 a contest was organized in order to investigate the problem of fusing visual and textual retrieval. Information fusion is a widely used technique to combine information from various sources to improve the performance of information retrieval. Fusion improvement relies on the assumption that the heterogeneity of multiple information sources allows self–correction of some errors leading to better results [3]. Medical documents often contain visual information as well as textual information and both are important for information retrieval [12]. The ImageCLEF benchmark addresses this problem and has organized a medical image retrieval task since 2004 [6]. So far it was observed in ImageCLEF that text–based systems strongly outperformed visual systems, sometimes by up to a factor of ten [11]. It is important to determine optimal fusion strategies allowing overall performance improvement as in the past some groups had combinations leading to poorer results than textual retrieval alone. The ImageCLEF@ICPR fusion task described in this paper is organized to address this goal, making available the four best visual and the four best textual runs of ImageCLEF 2009 including runs of various participating groups.

Information fusion, which originally comes from multi–sensor processing [17], can be classified by 3 fusion levels: signal level, feature level, and decision level [13] (also named the raw data level, representation level, and classifier level in [2]). The ImageCLEF@ICPR fusion task focuses on the decision level fusion, so the combination of the outputs of various systems [7]. Many fusion strategies have been proposed in the past. Using the maximum combination (*combMAX*), the sum combination (*combSUM*) and the multiplication of the sum and the number of non–zero scores(*combMNZ*) were proposed by [5] and are described in Section 2.3. Ideas such as Borda–fuse [1] and Condorcet–fuse [10] were rather inspired from voting systems. Borda–fuse consists of voting with a linear penalization based on the rank whereas Condorcet–fuse is based on pair–wise comparisons. Others strategies exist such as Markov models [4] and probability aggregation [9]. A terminology superposition also exists. For example, the round–robin strategy as analyzed in [17] is equivalent to the *combMAX* strategy, the Borda–fuse strategy, despite the idea being inspired from voting, is

in fact the *combSUM* strategy with descending weights for ranks. First proposed in 1994, *combMAX*, *combSUM*, and *combMNZ* are still the most frequently used fusion strategies and were taken as the base of our study. However, these three methods have limitations. On the one hand, *combMAX* favors the documents highly ranked in one system (*Dark Horse Effect* [14]) and is thus not robust to errors. On the other hand, *combSUM* and *combMNZ* favor the documents widely returned to minimize the errors (*Chorus Effect*) but relevant documents can obtain high ranks when they are returned by few systems. In this paper, we investigate a trade–off between these methods while using the sum of n maximums: *combSUM(*n*)MAX*.

Two other important issues of information fusion are the normalization of the input scores [8, 16] and the tuning of the respective weights (i.e contribution) given to each system [14, 15]. The normalization method proposed by Lee [8] consists of mapping the score to [0;1]. It was declared to perform best in [16]. Our study reused this normalization method, which is based on a topic basis or a run basis to produce normalized scores.

2. Methods

2.1 Dataset

The test data of the ImageCLEF@ICPR fusion task consist of 8 runs submitted to the ImageCLEF 2009 medical image retrieval task. 4 runs of the best textual retrieval systems and 4 representing the best visual systems were made available [1]. There are 25 query topics in ImageCLEFmed 2009. For each topic, a maximum of 1000 images can be present in a run. The format of each run follows the requirements of trec_eval.

Ranks and scores of the 8 runs are available as well as the ground truth for evaluation. Training data consists of the 4 best textual runs and the 3 best visual runs for the same task in 2008. The 7 runs in the training data and the 8 runs in the test data were not produced by the same systems. Therefore, weight selection on a run basis can not be applied to the test data.

2.2 Rank penalty vs. score normalization

To enable the combination of heterogeneous data, each image must be mapped to a value V (e.g. score, rank) that is normalized among all systems. Symbols employed are $\overline{V}$ for normalized values, and S and R for scores and ranks given by the input system.

[1] For more details about the retrieval systems, please visit ImageCLEF working notes available at `http://www.clef-campaign.org/`.

The scores given by the input systems are not homogeneous and require normalization. The normalization method proposed by Lee [8] is used:

$$\overline{V(S)} = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (1)$$

with S_{max} and S_{min} the highest and lowest score found. Two groups of normalized values were produced by either applying this method on run or topic basis: $\overline{V_{run}(S)}$ and $\overline{V_{topic}(S)}$.

The rank is always between 1 and 1000. However, low ranks need to be penalized as less relevant. Linear normalized rank values are obtained:

$$\overline{V_{linear}(R)} = N_{images} - R, \quad (2)$$

where N_{images} equals the lowest rank (1000 in our case). Experiments have shown that for most information retrieval systems, performance tends to decreases in a logarithmic manner [14]. As a consequence a logarithmic penalization function was tried:

$$\overline{V_{log}(R)} = \ln N_{images} - \ln R, \quad (3)$$

In the rest of the paper, $\overline{V}$ generally refers to one of these four groups of normalized values: $\overline{V_{topic}(S)}$, $\overline{V_{run}(S)}$, $\overline{V_{linear}(R)}$ and $\overline{V_{log}(R)}$.

2.3 Combination rules

combMAX computes the value for a result image i as the maximum value obtained over all N_k runs:

$$V_{\texttt{combMAX}}(i) = \arg \max_{k=1:N_k} (\overline{V_k(i)}). \quad (4)$$

combSUM computes the associated value of the image i as the sum of the $\overline{V}(i)$ over all N_k runs:

$$V_{\texttt{combSUM}}(i) = \sum_{k=1}^{N_k} \overline{V_k(i)}. \quad (5)$$

combMNZ aims at giving more importance to the documents retrieved by several systems:

$$V_{\texttt{combMNZ}}(i) = F(i) \sum_{k=1}^{N_k} \overline{V_k(i)}, \quad (6)$$

where $F(i)$ is the frequency of an image, counting the number of runs that retrieved the image i. Images that obtain identical values were arbitrarily ordered. The *combMAX* and *combSUM* rules both have drawbacks. *CombMAX* is not robust to errors as it is based one a single run for each image. *CombSUM* has the disadvantage of being based on all runs and thus includes runs

with low performance. As a trade–off the sum of N_{max} maximums rule combSUM(n)MAX is proposed:

$$V_{\texttt{combSUM(}n\texttt{)MAX}}(i) = \sum_{j=1}^{n} \arg \max_{k \in \mathcal{E}_{N_k} \setminus \mathcal{E}_j} (\overline{V_k(i)}), \quad (7)$$

with n the number of maximums to be summed and $\mathcal{E}_{N_k} \setminus \mathcal{E}_j$ the ensemble of N_k runs minus the j runs with maximum value for the image i. When $n = 1$, only 1 maximum is taken, which is equivalent to combMAX. Summing $n > 1$ maximums increases the stability of combMAX. When $n = N_k$, this strategy sums up all maximums and is equivalent to combSUM. $n < N_k$ can potentially avoid runs with low performance if assuming that runs with maximum scores or ranks have higher confidence and thus allow best retrieval performance.

As *combMNZ* proved to perform well, integrating the frequency is expected to improve performance. Instead of using a multiplication between the sum of values and the frequency, images are separated into pairs $\{F(i) : V(i)\}$ and are sorted hierarchically. Images with high frequency are ranked higher.

3. Results

An analysis was performed to analyze the distribution of the relevant documents in the training data. Each run in the training data contains 30 topics. Within each topic there are 1000 ranked images per topic. The 1000 ranks were divided into 100 intervals, and the number of relevant images were counted in each interval. As some topics contain few relevant images, all 30 topics were summed to obtain a more stable curve. Two curves containing the average numbers for all visual systems as well as all textual systems are shown in Figure 1.

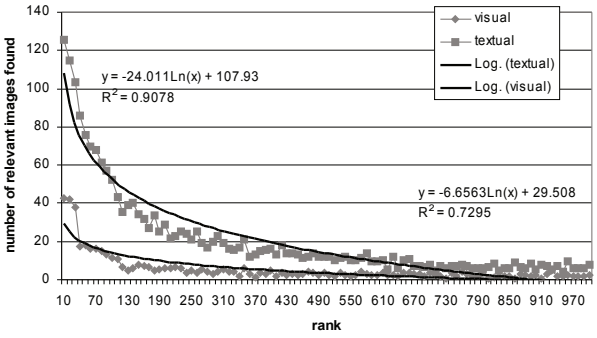

Figure 1. Distribution of the relevant documents in the training data.

In Table 1, the performance of the best fused runs are compared with the best runs of ImageCLEF 2009. The retrieval performance is measured using the mean average precision (MAP). MAPs obtained with various combination methods are shown in Figure 2.

Table 1. Best original vs. fused runs.

Run	2008	2009
best original textual run	0.2881	0.4293
best original visual run	0.0421	0.0136
best textual fusion run	0.3611	0.4766
best visual fusion run	0.0611	0.0198
best mixed fusion	0.3654	0.488

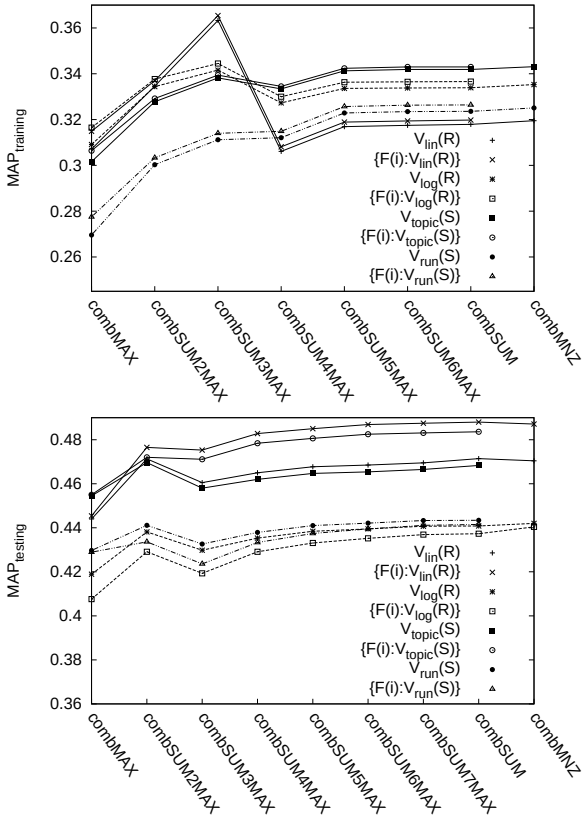

Figure 2. MAP on training and test data.

4. Interpretation

Two trends using logarithmic regression were calculated to analyze the distribution of relevant documents in Figure 1. Two observations can be made: 1) for both modalities the number of relevant images decreases logarithmically ($R^2 > 0.7$), which confirms Vogt [14]; 2) the quality of text retrieval is constantly 4 times better than that of visual. Fixed weights w were applied to

combine the two modalities with $w = 0.8$ for text systems and $w = 0.2$ for visual systems, and results are shown in Table 1. Two observations highlight the benefits of heterogeneity: 1) for both modalities best fused runs outperform all original runs; 2) multi–modal fusion outperformed the best run obtained with single modality fusion. Two–tailed paired t tests were performed in order to study the statistical significance of the two observations. Observation 2) is significant with both training ($p_{train} < 0.012$) and test ($p_{test} < 0.0116$) data. Observation 1) is significant with test ($p_{test} < 0.0243$) but not training ($p_{train} < 0.4149$) data.

The comparative analysis of *combMAX*, *combSUM* and *combMNZ* as well as *combSUM(*n*)MAX* is shown in Figure 2. Operators based on few maximums (left side of the graph: *combMAX*, *combSUM(*n*)MAX* with small values of n favor the *dark horse effect* whereas those based on several runs (*combSUM*, *combMNZ*) favor the *chorus effect*. The presence of coincident local minimum MAP for all techniques is due to the absence of both mentioned effects. For training data, maximum MAP was obtained with linear rank penalization using *combSUM3MAX* whereas for test data, log rank penalization using *combSUM* gave the best results. With logarithmic rank penalization, the behavior of MAP is the most stable among all techniques. This is in accordance with the descriptive analysis of the data where the relevance of images decreases with $\ln(R)$ (Figure 1). The performance using normalized score for fusion depends highly on score definition of each run.

In this paper, we studied the fusion of textual and visual retrieval systems. Fused runs outperform the original runs and combining visual information can significantly improve fusion performance. The logarithmic rank penalization is the most stable normalization strategy. As *Dark Horse Effect* oriented operators, *combSUM(*n*)MAX* outperforms *combMAX*, whereas *combSUM* and *combMNZ* give often close results on favoring *Chorus Effect*. In our experiments, no significant differences of performances were observed between the two effects and neither *Chorus* or *Dark Horse Effect* can be declared best. The improvement depends on the nature of the data to be fused.

References

[1] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284, New York, NY, USA, 2001. ACM.

[2] W. B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. Springer US, 2000.

[3] T. G. Dietterich. Ensemble methods in machine learning. In *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, London, UK, June 2000. Springer-Verlag.

[4] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001.

[5] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Text REtrieval Conference*, pages 243–252, 1993.

[6] W. Hersh, H. Müller, J. Kalpathy-Cramer, E. Kim, and X. Zhou. The consolidated ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*, 22(6):648–655, 2009.

[7] J. Kludas, E. Bruno, and S. Marchand-Maillet. Information fusion in multimedia information retrieval. In *Proceedings of 5th international Workshop on Adaptive Multimedia Retrieval (AMR)*, volume 4918, pages 147–159, Paris, France, June 2008. ACM.

[8] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM.

[9] D. Lillis, F. Toolan, R. Collier, and J. Dunnion. Probfuse: a probabilistic approach to data fusion. In *SIGIR '06: Proceedings of the 29th ACM SIGIR conference on Research and Development in information retrieval*, pages 139–146, New York, USA, 2006.

[10] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, New York, NY, USA, 2002. ACM.

[11] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, R. Said, B. Bakke, C. E. Kahn Jr., and W. Hersh. Overview of the CLEF 2009 medical image retrieval track. In *Working Notes of CLEF 2009*, Corfu, Greece, September 2009.

[12] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medicine–clinical benefits and future directions. *IJMI*, 73(1):1–23, February 2004.

[13] L. Valet, G. Mauris, and P. Bolon. A statistical overview of recent literature in information fusion. *Aerospace and Electronic Systems Magazine, IEEE*, 16(3):7–14, 2001.

[14] C. C. Vogt and G. W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, October 1999.

[15] S. Wu, Y. Bi, X. Zeng, and L. Han. Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Information Processing & Management*, 45(4):413–426, July 2009.

[16] S. Wu, F. Crestani, and Y. Bi. Evaluating score normalization methods in data fusion. In *Information Retrieval Technology, AIRS 2006*, pages 642–648, 2006.

[17] S. Wu and S. Mcclean. Performance prediction of data fusion for information retrieval. *Information Processing & Management*, 42(4):899–915, July 2006.