Fusion vs. Two-Stage for Multimodal Retrieval

Avi Arampatzis, Konstantinos Zagoris, and Savvas A. Chatzichristofis

Department of Electrical and Computer Engineering,
Democritus University of Thrace, Xanthi 67100, Greece
{avi,kzagoris,schatzic}@ee.duth.gr

Abstract. We compare two methods for retrieval from multimodal collections. The first is a score-based fusion of results, retrieved visually and textually. The second is a two-stage method that visually re-ranks the top-K results textually retrieved. We discuss their underlying hypotheses and practical limitations, and contact a comparative evaluation on a standardized snapshot of Wikipedia. Both methods are found to be significantly more effective than single-modality baselines, with no clear winner but with different robustness features. Nevertheless, two-stage retrieval provides efficiency benefits over fusion.

1 Introduction

Nowadays, information collections are not only large, but they may also be *multimodal*. Take as an example Wikipedia, where a single topic may be covered in several languages and include non-textual media such as image, sound, and video. Moreover, non-textual media may in turn be annotated.

We focus on two modalities, text and image. On the one hand, textual descriptions are key to retrieving relevant results for a topic, but at the same time provide little information about image content [5]. On the other hand, the visual content of images contains large amounts of information, which can hardly be described by words, making content-based image retrieval (CBIR) ineffective and computationally heavy in comparison to text retrieval. Thus, hybrid techniques which combine both worlds are becoming popular.

Traditionally, the method that has been followed in order to deal with multimodal databases is to search the modalities separately and fuse their results [4], e.g. with a linear combination of retrieval scores of all modalities per item. While fusion has been proven robust, we argue that it has a couple of issues: a) appropriate weighing of modalities and score normalization/combination are not trivial problems and may require training data, and b) if results are assessed by visual similarity only, fusion is not a theoretically sound method: the influence of textual scores may have a negative impact on the visual relevance of end-results.

An approach that may tackle the issues of fusion would be to search in a two-stage fashion: first rank with a secondary modality, draw a rank-threshold K, and then re-rank only the top-K items with the primary modality. The assumption on which such a two-stage setup is based on is the existence of a primary modality (i.e. the one targeted and assessed by users) and its success would largely depend on the relative effectiveness of the two modalities involved. For example, if in the top-K, text retrieval performs better

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 759–762, 2011.

than CBIR, then CBIR is redundant. Thus, the underlying hypothesis is that CBIR can do better than text retrieval in the top-K results retrieved by text.

Thresholding for two-stage retrieval can be performed statically (i.e. a fixed preselected threshold for all topics, e.g. [7]) or in a dynamic manner (i.e. a variable threshold optimizing a pre-defined measure per topic, e.g. [2]). In recent literature, the effectiveness of static thresholding has been mixed. For instance, static thresholding was found to perform worse in mean average precision (MAP) than the text-only with pseudo relevance feedback baseline in [7] (but better than fusing image and text modalities by a weighted-sum). However, others found that two-stage retrieval with dynamic thresholding is more effective and robust than static thresholding, performing significantly better than a text-only baseline [2].

A possible drawback of two-stage setups is that visually relevant images with empty or very noise text modalities would be completely missed, since they will not be retrieved by the first stage. Moreover, if there are any improvements compared to single-stage text-only or image-only setups, these will first show up on early precision since only the top results are re-ranked; MAP or other measures may improve as a side effect. Fusion does not have these problems.

Next, we provide an experimental comparison of fusion to two-stage retrieval. Although we argued theoretically against fusion, in view also of the underlying assumption, hypothesis and drawbacks of two-stage retrieval, a comparison of the effectiveness of the two methods is in order.

2 An Experiment on Wikipedia

In this section, we report on experiments performed on the ImageCLEF 2010 Wikipedia test collection, which consists of 237434 images associated with noisy and incomplete user-supplied textual annotations. There are 70 test topics, each one consisting of a textual and a visual part, with one or more example images. The topics were assessed by visual similarity to the image examples.

We index the images with two descriptors that capture global image features: the Joint Composite Descriptor (JCD) and the Spatial Color Distribution (SpCD) [3]. For text indexing and retrieval, we employ the Lemur Toolkit V4.11 and Indri V2.11 with the tf.idf retrieval model; tf.idf has been found to work well with the the dynamic thresholding method we will describe in Section 2.2 [1]. We use the default settings that come with these versions of the system except that we enable Krovetz stemming. We index only the English annotations, and use only the English query of the topics. We evaluate on the top-1000 results with MAP, precision at 10 and 20, and bpref.

2.1 Fusion of Modalities

Let i the index running over example images ($i = 1, 2, \ldots$) and j running over the visual descriptors ($j \in \{1, 2\}$). Thus, DESC_{ji} is the score of a collection item against the ith example image for the jth descriptor. We normalize DESC_{ji} values with MinMax, taking the maximum score seen across example images per descriptor. Assuming that the descriptors capture orthogonal information, we add their scores per example image. Then, to take into account all example images, the natural combination is to assign to

each collection image the maximum similarity seen from its comparisons to all example images; this can be interpreted as looking for images similar to *any* of the example images. Incorporating text, again as an orthogonal modality, we add its contribution. Summarizing, the score s for a collection image against the topic is defined as:

$$s = (1 - w) \max_i \left(\sum_j \mathrm{MinMax}(\mathrm{DESC}_{ji}) \right) + w \, \mathrm{MinMax}(\mathrm{tf.idf}) \,. \qquad (1)$$

The parameter w controls the relative contribution of the two media; for $w = 1$ retrieval is based only on text while for $w = 0$ is based only on image. We report for five w values between 0 and 1.

2.2 Dynamic Two-Stage Retrieval

For dynamic thresholding, we use the Score-Distributional Threshold Optimization (SDTO) as described in [1]. The SDTO method fits a binary mixture of probability distributions on the score distribution (SD). For tf.idf scores, we used the *technically truncated* model of a normal-exponential mixture. The method normalizes retrieval scores to probabilities of relevance (prels), enabling the optimization of K for any user-defined effectiveness measure. Per query, we search for the optimal K in $[0, 2500]$. Thus, for estimation with the SDTO we truncate at the score corresponding to rank 2500 but use no truncation at high scores as tf.idf has no theoretical maximum.

We experiment with the SDTO by thresholding on prel. This was found in [2] to be more effective and robust than thresholding on estimated precision. Thresholding on fixed prels happens to optimize *linear utility measures* [6]. We report for five prel thresholds. The top-K results are re-ranked using Equation 1 for $w = 0$.

2.3 Experimental Results

Table 1 presents the effectiveness of fusion and two-stage against text- and image-only runs. Irrespective of measure, the best parameter values are roughly at: 0.6666–0.8000

Table 1. Retrieval effectiveness for fusion and dynamic two-stage retrieval. The best results per measure and retrieval type are in boldface. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05 ($^{\triangle\triangledown}$), 0.01 ($^{\blacktriangle}$), and 0.001 ($^{\blacktriangle\blacktriangledown}$), against the text-only baseline.

		MAP	P@10	P@20	bpref
text-only		.1293	.3614	.3307	.1809
fusion w	.9000	.1380$^{\blacktriangle}$	.3786$^{\triangle}$	.3414$^{\triangle}$	.1901$^{\blacktriangle}$
	.8000	**.1410**$^{\blacktriangle}$	.4029$^{\blacktriangle}$	.3514$^{\blacktriangle}$	.1955$^{\blacktriangle}$
	.6666	.1403$^{\blacktriangle}$	.4129$^{\blacktriangle}$	**.3664**$^{\blacktriangle}$	**.1969**$^{\blacktriangle}$
	.5000	.1185^{-}	**.4157**$^{\triangle}$	.3657$^{\triangle}$	.1758^{-}
	.3333	.0767$^{\blacktriangledown}$	.3871^{-}	.3329^{-}	.1278$^{\blacktriangledown}$
two-stage θ	.9900	.1376$^{\triangle}$	.4286$^{\triangle}$	.3714$^{\triangle}$	.1899$^{\triangle}$
	.9500	.1390$^{\triangle}$	.4314$^{\triangle}$	.3771$^{\triangle}$	.1917$^{\triangle}$
	.8000	**.1428**$^{\triangle}$	**.4443**$^{\triangle}$	**.3857**$^{\triangle}$	**.1959**$^{\triangle}$
	.5000	.1405$^{\triangle}$	.4357$^{\triangle}$	.3821$^{\triangle}$	.1943$^{\triangle}$
	.3333	.1403$^{\triangle}$	.4357$^{\triangle}$	.3807$^{\triangle}$	.1942$^{\triangle}$
image-only		.0107$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0402$^{\blacktriangledown}$

for fusion's w, and 0.8000 for two-stage's θ. Both methods perform significantly better than text-only and far better than image-only. On the one hand, two-stage achieves better results than fusion, but it has more variability across topics: fusion passes the test at lower significance levels (i.e. higher confidence). On the other hand, effectiveness is less sensitive to the values of θ than the values of w: two-stage provides significant improvements in all measures for a wide range of thresholds (i.e. 0.3333–0.9900), while fusion can significantly deteriorate effectiveness for unsuitable choices of w.

3 Conclusions

We compared fusion to two-stage retrieval from multimodal databases and found that both methods are significantly better that text- and image-only baselines. Indicatively, the largest improvements in MAP against the text-only baseline are +9.0% and +10.4% for fusion and two-stage respectively, while the corresponding improvements in P@10 are +15.0% and +22.9%.

While two-stage performs better than fusion in 3 out of 4 measures, improvements are statistically non-significant at the 0.05 level. Further, both methods are robust in different ways: fusion provides less variability across topics but it is sensitive to the weighing parameter of the contributing media, while two-stage provides a much lower sensitivity to its thresholding parameter but has a higher variability. Nevertheless, two-stage has an obvious efficiency benefit over fusion: it cuts down greatly on costly image operations. Although we have not measured running times, only the 0.02–0.05% of the items (on average) had to be scored at the image stage. While there is some overhead for estimating thresholds, this offsets only a small part of the efficiency gains.

References

1. Arampatzis, A., Kamps, J., Robertson, S.: Where to stop reading a ranked list: threshold optimization using truncated score distributions. In: SIGIR, pp. 524–531. ACM, New York (2009)
2. Arampatzis, A., Zagoris, K., Chatzichristofis, S.A.: Dynamic two-stage image retrieval from large multimodal databases. In: ECIR 2011. LNCS, vol. 6611. Springer, Heidelberg (2011)
3. Chatzichristofis, S.A., Arampatzis, A.: Late fusion of compact composite descriptors for retrieval from heterogeneous image databases. In: SIGIR, pp. 825–826. ACM, New York (2010)
4. Depeursinge, A., Muller, H.: Fusion techniques for combining textual and visual information retrieval. In: ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Springer, Heidelberg (2010)
5. van Leuken, R.H., Pueyo, L.G., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: WWW, pp. 341–350. ACM, New York (2009)
6. Lewis, D.D.: Evaluating and optimizing autonomous text classification systems. In: SIGIR, pp. 246–254. ACM Press, New York (1995)
7. Maillot, N., Chevallet, J.-P., Lim, J.-H.: Inter-media pseudo-relevance feedback application to imageCLEF 2006 photo retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 735–738. Springer, Heidelberg (2007)