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.