Text-Based Clustering of the ImageCLEFphoto Collection for Augmenting the Retrieved Results

Osama El Demerdash, Leila Kosseim, and Sabine Bergler

CLaC Laboratory - Department of Computer Science & Software Engineering - Concordia University
{osama_el,kosseim,bergler}@cse.concordia.ca

Abstract. We present our participation in the 2007 ImageCLEF photographic ad-hoc retrieval task. Our first participation in this year's ImageCLEF comprised six runs. The main purpose of three of these runs was to evaluate the text and visual retrieval tools as well as their combination in the context of the given task. The other purpose of our participation was to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of the ImageCLEF Ad hoc task. We used the preclustered collection to augment the search results of the retrieval engines. For retrieval, we used two publicly available libraries; *Apache Lucene* for text and *LIRE* for visual retrieval. The clustered-augmented results reduced slightly the precision of the initial runs. While the aspired results have not yet been achieved, we note that the task is useful in assessing the validity of the clusters.

1 Introduction

In this paper, we present our participation in the 2007 ImageCLEF photographic ad-hoc retrieval task. The task deals with answering 60 queries of variable complexity from a repository of 20,000 photographic images in the IAPR TC-12 collection. A full description of the task and the collection can be found in [1]. Our first participation in this year's ImageCLEF comprised six runs. The main purpose of three of these runs was to evaluate the text and content-based retrieval tools in the context of the given task. We therefore would like to stress that the evaluation of these tools can only be considered under the given parameters of the task, including the queries, the image collection and our utilization of these tools.

The other purpose of our participation was to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of the ImageCLEF Ad hoc retrieval task. While this task of ImageCLEF was not intended for the evaluation of interactive methods, it could still be useful in the evaluation of certain aspects of such methods such as the validity of the initial clusters in our case.

2 Related Work

Clustering, as an unsupervised machine learning mechanism, has rarely been investigated within the context of the ImageCLEF ad-hoc Retrieval task. This

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 562–568, 2008.