Chapter 3
Creating Realistic Topics for Image Retrieval Evaluation

Henning Müller

Abstract This chapter describes the various ways for creating realistic query topics in the context of image retrieval evaluation campaigns such as ImageCLEF. A short overview describes general ways of creating topics, from complete laboratory style evaluations based on the technical capabilities of systems to real-world applications with real end users. The chapter offers help to those planning to evaluate systems on how to develop challenging and realistic topics based on knowledge of the users and of the capabilities of systems. Information sources for created topics are detailed. The main analysis will be the ImageCLEF tasks, and especially the medical retrieval tasks, where many different ways for creating topics have been analyzed over the years.

3.1 Introduction

Evaluation has always been an important aspect of systems development and demonstrating technical progress in all fields of research, including information retrieval. Creating formalised statements of user’s information needs (topics) is a core part of IR evaluation using test collections. Topics are used to compare techniques in a particular field of research; however, creating realistic and effective topics is far from trivial. In information retrieval, the first systematic evaluation of research systems were the Cranfield tests in 1962 (Cleverdon, 1962). These tests mention the following as requirements for evaluation: the existence of a data set; the creation of query tasks and detailed topics that correspond to a user’s information need; and a judgement of relevance for all documents/images in the collection with respect to the created topics. Almost all current evaluation campaigns such as TREC and

Business Information Systems, University of Applied Sciences Western Switzerland (HES–SO), TechnoArk 3, 3960 Sierre, Switzerland, e-mail: henning.mueller@hevs.ch

CLEF\textsuperscript{2} are still based on this paradigm (Harman, 1992; Savoy, 2002), although with increasing database size judging all items in a database for relevance is not possible and pooling is usually used to limit the amount of work required for the judgments (Sparck Jones and van Rijsbergen, 1975). (See Chapter 4 for more details regarding relevance assessments.) Thus topic creation has been an integral part of the evaluation process in information retrieval.

This chapter focuses on the evaluation of image retrieval, however, rather than textual information retrieval. Image retrieval has been a very active domain over the past 25 years (Smeulders et al, 2000) but evaluation of image retrieval has rather been neglected (Müller et al, 2001) over much of this period. Over the last ten years, this has slowly changed and a large number of evaluation campaigns and more systematic evaluation approaches have also started in visual information retrieval. After initial proposals from Gunther and Beretta (2001) with general ideas, TRECVis\textsuperscript{3} has been the first campaign to systematically evaluate video retrieval from large-scale archives with news footage (Smeaton et al, 2003). Other campaigns more focused on image retrieval, such as ImageCLEF\textsuperscript{4} or ImageEval\textsuperscript{5}, followed only a little later.

In terms of topic creation, only very limited systematic analysis has taken place and one of the few papers really describing the process of topic generation for ImageCLEF is by Grubinger and Clough (2007). For most other evaluation campaigns, available data sources such as user log files have been used from a variety of sources such as Web log files (Müller et al, 2007), or library log files (Clough et al, 2006). Another approach is to integrate the participants into the creation of topics (Tsikrika and Kludas, 2009). The goal of topic development is usually to create topics that:

- correspond to a specific user model, i.e. a person searching for information in a particular context;
- correspond to real needs of operational image retrieval systems;
- are at least partly solvable with the existing technology;
- are diverse to allow a good part of the retrieval functionality to be tested and a large part of the data set to be explored;
- differ in coverage from rather broad to very specific needs;
- are solvable with documents from the given collection.

Another problem when considering analyzing visual information retrieval is how to express the information need of a potential user precisely. Information needs can generally be described in words, but for topic generation they can be represented with either text or visual examples, which determines which types of system can be evaluated. Most often, text is used for expressing the topic and textual information retrieval is much further advanced than visual retrieval in this respect. If the goal

---

\textsuperscript{2} Cross Language Evaluation Forum, \url{http://www.clef-campaign.org/}
\textsuperscript{3} \url{http://trecvid.nist.gov/}
\textsuperscript{4} \url{http://www.imageclef.org/}
\textsuperscript{5} \url{http://www.imageval.org/}