Benchmarking Image and Video Retrieval: an Overview

Stéphane Marchand-Maillet
Viper group for Multimedia Retrieval
University of Geneva - 24, Rue Général Dufour
CH-1211 Geneva – Switzerland
http://viper.unige.ch

Marcel Worring
Intelligent Sensory Information Systems
University of Amsterdam - Kruislaan 403
1098 SJ Amsterdam - The Netherlands
worring@science.uva.nl

ABSTRACT

Multimedia Information Retrieval (IR) techniques and associated systems are now numerous and justify the development of strategies and actions to objectively evaluate their capabilities. A number of initiatives following this line exist, each with its own peculiarities. In this paper, we take a bird's eye view on benchmarking multimedia IR systems (with the particular case of image and video retrieval) and summarize contributions made to a dedicated special session at the ACM Multimedia Information Retrieval Workshop (ACM MIR2006).

From the analysis of a classical IR system, we identify locations of interest for evaluation of performance. We review proposals made in the context of existing benchmarks, each specialized in its own aspect and media.

Categories and Subject Descriptors

H.3.3 [**Information search and retrieval**]: *Search process*; H.3.1 [**Content Analysis and Indexing**]: *Indexing methods*; H.5.1 [**Multimedia Information systems**]: *Evaluation/Methodology*

Keywords

Evaluation, multimedia, information retrieval, image, video

1. INTRODUCTION

Multimedia Information Retrieval (IR) techniques and associated systems are now numerous and justify the development of strategies and actions to evaluate their capabilities. A number of initiatives following this line exist, each with its own peculiarities.

Even if the most recent techniques for multimedia IR have drastically departed from their origin, the link to classical (text) IR is still present. Most evaluation consortium present their initiatives in relation to the TREC enterprise.

The Text REtrieval Conference (TREC for short) [14] is an annual series of evaluation campaigns for IR systems. TREC follows the Cranfield laboratory setup [4] for IR system evaluation, which is based on the abstraction of a test collection comprising a set of *documents*, a set of information needs (called *topics*) and a set of *relevance judgments* associating documents to topics with a binary relevance relationship.

In this paper, we take a bird's eye view on benchmarking multimedia IR systems. When required, we specialize in image and video IR systems. For the case of text IR evaluation, the reader is directly referred to TREC [14] and for audio IR evaluation, the reader is referred to MIREX [10] and IMIRSEL [7]. We wish to look at the construction of a benchmark from a global perspective. Essentially, we wish to make sure that the evaluation platform constructed will help generating results that are relevant to the advancement of the research field in question.

In section 2, we look at what aspects of IR can be assessed and how they should be assessed. This discussion is supported by contributions made to a dedicated special session at the ACM Multimedia Information Retrieval Workshop (ACM MIR2006). While doing so, we list adjunct performance measures that have been proposed and how they are actually used. In section 3, we look at existing benchmarks and discuss their features according to our context.

2. EVALUATING IR SYSTEMS

IR systems and protocols are generally complex, composed of parts running underlying technologies attached to various domains of competence. From the sketch of a classical IR system, we identify possible variables which may play a role in the *overall performance* of the system or, equivalently, the *satisfaction* it provides to any of its users. In [8], this is looked at from the even wider perspective of the capability of a system to "advance the primary endeavor within which the primary IR episode is embedded, [such as] work learning or entertainment". We assume that whatever the context of the IR episode, an *information need* is to be satisfied and is materialized by the *query* formulated to the system (the "overall information seeking task" according to [13]).

Figure 1 shows the structure of a classical IR setup and we mark points where evaluation may be relevant. These points are detailed in the sections below (see related numbering), in relation to what benchmarks usually propose. The aim is to show that these points actually represent the "degrees of freedom" of any IR system and that any of them may affect the overall performance. We therefore assert that all issues should ideally be taken into account for a complete capture

MIR'06, October 26–27, 2006, Santa Barbara, California, USA.

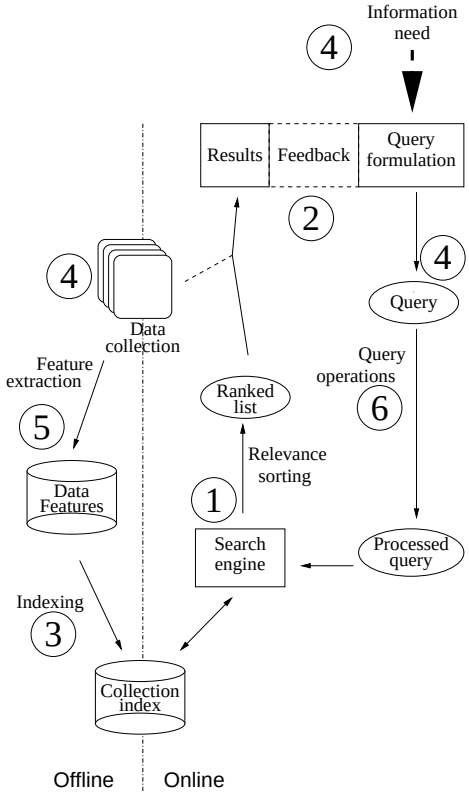

Figure 1: Structure of a classical IR system and points where evaluation may be relevant

of the performance.

2.1 Retrieval performance

Following the philosophy of the Cranfield experiment and hence TREC, most IR benchmarks (multimedia or not) focus on evaluating retrieval performance. That is, they naturally concentrate on the technology underlying the main part of the system, the search engine. Hence, performance is related to the ability of the system to respond to the query by a list of documents where the document judged as being most relevant are ranked first. The protocol for evaluation in this case will be the simulation of a real user creating automated (positive and negative) relevance from given relevance judgments in an iterative query loop.

Related measures include for the most the notion of *Precision* (P) and *Recall* (R), often materialized as P-R curves [1] . These curves may be shown at various steps of the complete IR episode to *e.g.* demonstrate the improvement of the system performance from the use of relevance feedback.

Initial relevance feedback includes large classes of items relevant to the current information need. Hence, when generating automated feedback, choices are to be made on which documents to select as feedback. This is further made necessary since the simulation of the user behavior leads to providing the system with only a realistic amount of feedback (typically tens of documents). This results in a possible variation of the system performance, depending on the choices made. To provide a legible overview, average P-R curves are often used. To further summarize the performance of a system, Mean Average Precision (MAP), which typically represents the area below the average P-R curve is also often used [13, 15].

Also worth noting is the binary preference measure *bpref* proposed by major TREC actors in [3] and discussed in [15]. The reader is referred to [12] for further details on retrieval performance measures. An interesting study on how these measures may be found to be incomplete is given in [11].

We finally point out the fact that a number of modern multimedia IR techniques often derive from a learning-based context, close to classification tasks where ROC curves may be used to illustrate system performance. The link here is that ROC parameters (*specificity* and *sensitivity*) are of the same essence than the notions of Precision and Recall.

2.2 User interaction

The above measurements apply to any type of data. However, [8] demonstrates that in general (and in particular in the case of video retrieval), the fact that there is a human in the retrieval loop introduces a new variability.

Clearly, since relevance feedback has to be provided to the system, there is the need for creating a user interface (in the broad sense) to the system. This interface generally includes a result display, which may be vary in the way the ranked list is presented, but also calls for the definition of a representation of the data. It is often from this very presentation that feedback is to be provided to the system.

While Web search engines have accustomed users to presentations of documents excerpts, there is no clear such scenario for the case of video. In [8], the notion and effectivity of *video surrogates* is thoroughly discussed in relation to main search tasks with the aim of [creating] "video surrogates that enhance the search experience". Here, search is understood in the context of *lookup*, *learning* and *investigation*, which form the human-centered tasks. The measures related to the user "distinguish physical, cognitive and attitudinal characteristics".

This specific study concentrates on how to present complex data to the user for efficient usage. This is clearly related to and may be completed by generic interface usability studies including systematic user tests such as that described in [2, 16].

2.3 Indexing

Combining the above two sections leads to a system where the user would be able to communicate optimally (*e.g.*, without any loss of information, with no ambiguity) with a core engine that shows perfect performance (*e.g.*, perfect Precision, whatever the Recall level). References [5, 8, 13, 15] and associated benchmarks primarily gather these indications. This however does no clarify how the internals of a given system would perform. Following the idea of Unit Testing, one may detail a benchmark that digs into the system nuts and bolts to determine unitary performances.

In this part, we are particularly concerned with the way data are accessed and indexed. The evaluation in this section is looking at notions such as:

- *size*: what will be the cost in terms of storage capacity for obtaining a fluent access to the original data? It is known that working with video is cumbersome in terms of the volume of original data and one quickly finds out that the size of the indices generated becomes a problem at least equally important;

- *time*: adjunct to the above is the problem of the time efficiency of the indexing strategy. As a relation to the above evaluation for usability (section 2.2), one wishes to satisfy constraints where the delay between query expression and result display is as short as seconds. One bottleneck in this process is possibly the access to the data via the index;
- *scalability*: While responding to the above question using a given setup (documents, topics and relevance), it is also important to evaluate the capacity of the system to scale up to other setups where any of the parameters (*e.g.*, corpus size, number of queries) is varying.

Much of these evaluations are classically done in the field of DBMS testing. Some parts of associate standard benchmarks could readily be transposed in the field of (multimedia) IR.

2.4 Benchmark input

So far, the investigation has focused on the system performance and user satisfaction. An important aspect to look at is the validity of the context in which the evaluation is performed. Here, this is represented by the base setup we are working with, the available data collection (the documents), the queries (topics) and the ground truth (relevance).

2.4.1 Data (documents)

The data against which the system is benchmarked clearly introduces a bias in the evaluation. Ideally, the data should be chosen in terms of its *variability* and *volume* so as to span and cover the complete domain in question. One straightforward example criteria of this coverage is simply that if the data volume would grow, its variability would not changed. In other words, introducing new data would not introduce any new aspect of the domain.

Another important aspect associated to the data is the fact that it should create a favorable statistical context for one to demonstrate the statistical significance of any variation of results from one system to another[1]. This is not always the case and interesting insights and models for this aspect are given in [6].

The issue of providing data to benchmarks is discussed practically in [13] and opposed to the crude reality leading to the fact that any benchmark should be happy with the data it may gather. Problems generally related to ownership and copyright often strongly block the process of developing ideal corpora for benchmarks.

The current trend however seems to be that the Web is used as source for data provided its "collective ownership". Classical such sources include Flickr [9] and Wikipedia [15].

2.4.2 Queries (topics)

The discussion readily applies to the set of queries (topics) that the benchmark is using as base for the assessment. The criteria of volume (number) and diversity (variability) are valid here and should also create a favorable context for demonstrating statistical robustness of the results published.

Queries are also supposed to represent what a real user of the system would instantiate as usage scenario. Hence, it may be important that the applicative relevance of the query set is assessed [5]. The way to go about such an evaluation is based on (industrial) domain experts or available market studies (if any). An alternative is to set up mock systems and collect and analyze usage logs for deriving such parameters.

[1]Here, we instantiate a new system as soon as any change in the testing strategy or tested technique is introduced (new testing parameter, new testing data, ...)

2.4.3 Ground truth (judgments, annotation)

A complementary aspect is the base knowledge with and against which the system will be evaluated. From the Cranfield setup and for the core engine evaluation, this is given in an *implicit* manner under the form of relevance judgments. Each document is tagged by (possibly several) experts with a degree of relevance to each query (topic). This degree is generally rather binary (relevant, irrelevant) added with a neutral (no opinion) state.

A rather more advanced way of providing this ground truth is by *explicit* annotation. This subject clearly extends beyond the scope of this summary paper. We simply note that annotation may be used to generate relevance judgments. The goal is to simulate users' reaction to result and generate automated feedback, which may account for a global context rather than be simply item-based. Annotation is also the methid of choice for classifation data [13].

In any case, the quality of the ground truth should be assessed. At the very least, any potential corruption of results from noise within that data should be modeled. In that case, obtaining multiple judgments from experts and using average P-R results is one way to smooth out flaws, subjectivity or simply ambiguity from within the ground truth.

2.4.4 Users

Users are not considered as a benchmark input but their base knowledge (class) may be. An IR episode will occur in a completely different manner if the user performing the test is a domain expert, a generic user or a computer scientist. The ability or expectation of each class of users may drastically bias the final assessment of the quality of this search episode.

2.5 Data representation

Two views may be taken over the problem of data representation. From a multimedia processing point of view, data representation represents the *feature extraction* process, the process of transforming the data into a machine understandable representation, from which the index will essentially be constructed. The basic features to be extracted typically derive from low-level processing with increasing complexity. From color, texture, frequency, one may move to regions, motion or pitch. These features being directly extracted from the raw signal, their evaluation rather concerns their ability to provide a useful representation of the data with useful properties mostly deriving from the notion of invariance. Benchmarks for these features exist in their respective domains.

Similarly, high-level (possibly multimodal) features such as face or object detection or recognition, OCR, speech transcription, boundary detection are associated with testbeds assessing the performance of these algorithms in several conditions and under chosen constraints. In the context of benchmarks, this may even become a task in itself (*e.g.* the high-level feature extraction task in TRECVid [13] or the object identification task in ImagEval [5]).

Another perspective on data representation that may be

taken and evaluated is the problem of processing the data to provide a useful representation to the user for *e.g.* feedback acquisition ("metadata akin to 'glosses' or 'abstracts' that 'stand for' the full [multimedia] object" [8]). This clearly relates to section 2.2 as evaluation will mainly be done in terms of user satisfaction.

Going a step further leads to the data representation moving from a passive display of results to an active way of browsing the complete collection or a support to determine the 'gist' of the data [8].

2.6 Query operations

The specific case of multimedia IR does not lend itself easily to query operations such as query expansions [1]. However, should this be the case, it would create a further operation whose performance impacts the overall performance. Thus requiring specific evaluation.

3. IMPLEMENTATION

Implementations of the various benchmark perspectives and scenarios described above exist and are thoroughly listed and discussed within [13], in complement to the specific discussions found in [5, 8, 15]. A related extra reference is the ImageCLEF effort [9]. These mostly address the issues listed in sections 2.1 and 2.2, while including some of the aspects discussed in section 2.4. Indexing benchmark (section 2.3) is a different domain overall as multimedia IR techniques often inherit from what is developed in DBMS research. Similarly, aspects developed in sections 2.5 and 2.6 are often considered as independent processes with their own base evaluation and in general not included in classical IR benchmarks.

4. CONCLUSION

Multimedia IR may be evaluated from several perspectives. Several benchmarks (as listed along the references cited here) are now mature and depart from the crude idea of a *competition* that may take place within communities. Rather, the aim is to highlight strengths and weaknesses (locations where room for improvment exists) of existing strategies or systems. Hence, it is important that these evaluation campaigns remain open to all aspects of what makes a IR system, as briefly sketched in this paper.

Existing benchmarks certainly contribute to the expansion of the field rather than promoting the "survival of the fittest". For many reasons, benchmarks are a necessity to structure a field of research and create a true community with an identity that then may be strong enough to resolve some issues that smaller entities would not be able to overcome. This may be at the cost of a constrained context but it is to the players in the field to open discussions and propose alternative solutions. In this respect, we certainly adhere to the discussion put forward in [13].

5. ACKNOWLEDGMENTS

The first author wish to acknowledge the support of the Swiss National Science Foundation for support under subside 2100-066648 and the NCCR (IM)2.

6. REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc., 1999.

[2] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[3] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR 2004*, New York, NY, USA, 2004.

[4] C. W. Cleverdon. The Cranfield tests ion index language devices. *ASLib proceedings*, 19(6):173–192, 1967.

[5] C. Fluhr, P-A Moëllic, and P. Hede. Usage-oriented multimedia information retrieval technological evaluation. In *Proceeding of the ACM Workshop on Multimedia Information Retrieval (MIR2006), special session on "Benchmarking Image and Video Retrieval Systems"*, Santa Barbara, CA, 2006.

[6] Cormack G.V. and Lynam T.R. Statistical precision of information retrieval evaluation. In *SIGIR 2006*, Seattle, USA, 2006.

[7] The International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) Project. `http://www.music-ir.org/evaluation`.

[8] G. Marchionini. Human performance measures for video retrieval. In *Proceeding of the ACM Workshop on Multimedia Information Retrieval (MIR2006), special session on "Benchmarking Image and Video Retrieval Systems"*, Santa Barbara, CA, 2006.

[9] ImageCLEF: the image retrieval track within CLEF. `http://ir.shef.ac.uk/imageclef/`.

[10] MIREX: Music Information Retrieval Evaluation eXchange. `http://www.music-ir.org/mirex2006`.

[11] Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about Corel – Evaluation in image retrieval. In *Proceedings of CIVR2002*, London, UK, July 2002.

[12] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters (Special Issue on Image and Video Indexing)*, 22(5):593–601, 2001. H. Bunke and X. Jiang Eds.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceeding of the ACM Workshop on Multimedia Information Retrieval (MIR2006), special session on "Benchmarking Image and Video Retrieval Systems"*, Santa Barbara, CA, 2006.

[14] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, 2005.

[15] T. Westerveld and R. W. Zwol. Benchmarking multimedia search in structured collections. In *Proceeding of the ACM Workshop on Multimedia Information Retrieval (MIR2006), special session on "Benchmarking Image and Video Retrieval Systems"*, Santa Barbara, CA, 2006.

[16] Ryen W. White, Joemon M. Jose, and Ian Ruthven. Adapting to evolving needs: Evaluating a behaviour-based search interface. In *17th Annual Human-Computer Interaction Conference (HCI 2003)*, New Orleans, U.S.A, 2003.