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[1] and

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

[1] Text REtrieval Conference, http://trec.nist.gov/

CLEF[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, TRECVid[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[4] or ImageEval[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 of a benchmark is to evaluate both visual and textual retrieval systems (and also

[2] Cross Language Evaluation Forum, `http://www.clef-campaign.org/`

[3] `http://trecvid.nist.gov/`

[4] `http://www.imageclef.org/`

[5] `http://www.imageval.org/`

combined retrieval), both media need to be represented in the query formulation. Whereas text can in this case easily be taken from usage log files, image examples are only very rarely available directly from such log files, as there are only very few visual systems in daily use. The choice of images for a query thus becomes an important part of the process and this is most often not analyzed further. Combined visual and textual retrieval really has the potential to improve current information access systems, but the results of evaluation campaigns to date also show how difficult these combinations are to work with.

In several evaluation tasks (Grubinger and Clough, 2007; Müller et al, 2009) the topics are classified into whether they mainly correspond to visual search tasks, where image analysis can be of use; to semantic search tasks, where mainly text retrieval can be useful; or to mixed tasks where the two can be expected to be useful. This classification is usually performed manually by experienced researchers and the results show that this classification is possible when being at least partly familiat with the database. This also means that systems could automatically determine the necessary resources for optimizing retrieval results if this knowledge can be formalized.

Another axis to take into account when developing topics is the topic difficulty, which needs to be challenging for existing systems employed and so rather difficult, but still correspond to the capabilities of the techniques. Particularly when pooling is used, the expected number of relevant images is also important as an excessively large number of relevant images can result in a large number of relevant documents remaining un–judged. On the other hand, a very small number of relevant documents can result in distorted performance measures if only one or two documents are relevant. Topic quantity is another important question that has been analyzed over many years. This is particularly important for getting stable/robust results and avoiding systems being ranked in a random order. Experiences in TREC suggest that at least 25 query topics are necessary for obtaining relatively stable results (Voorhees and Harmann, 2000), whereas others estimate this number to be much higher and near to 200–300 topics (Sparck Jones and van Rijsbergen, 1975). In general 25–50 query topics are recommended for relatively stable results.

An important link exists between the topic development and the relevance judgement process. TREC generally proposes that the topic creator should judge the relevant images themselves so the exact reasoning behind creating the topic can be taken into account for the judgment and means that this corresponds to one clear information need of a particular person. On the other hand, relevance of images has been shown to depend on the person, the situation and is not stable over time even for the same person. Thus, it was often proposed to have several judgments from different people so that the variability and subjectivity of the topics can be measured, e.g. using a kappa score (Müller et al, 2009). In general, results in ImageCLEF suggest that the judgments for image–based topics have less variation than for text–based query topics.

3.2 User Models and Information Sources

This section describes the underlying user models for image retrieval evaluation. Many purely image analysis benchmarks such as PASCAL[6] (Everingham et al, 2006) lack a concrete user model and involve rather basic scientific research tasks without any clearly visible application in mind. Examples for such topics can be detecting dogs or cats in images, which can then be used for future automatic annotation of images.

In general, when specific applications are identified, an appropriate user model is chosen such as journalists searching for images (Markkula and Sormunen, 1998) or Web users operating an image search engine (Goodrum, 2000). This can subsequently be taken into account for the definition of relevance in the evaluation. Relevance in itself is a rather poorly defined concept subject to much interpretation (Mizzaro, 1997) and having a clear user model and goal in mind can reduce this subjectivity. More on relevance judgments can be found in Chapter 4.

3.2.1 Machine–Oriented Evaluation

In image processing and many pattern recognition tasks involving images, the tasks for evaluation tools are more oriented towards advancing the current capabilities of techniques rather than towards real applications involving end users. This does not mean that these tasks cannot be useful, but care needs to be taken that tasks and databases are not too much oriented towards the capabilities of particular algorithms.

In the large majority of evaluation settings in image analysis, objects are to be detected in images such as in the PASCAL network of excellence (Everingham et al, 2006), or images are to be classified into a set of categories (Deselaers et al, 2007). This might currently not deliver results for real applications but it can be a preliminary step to developing tools that can subsequently help in such applications. Many other tasks have a user model in mind, such as clinicians searching for images but then use an outline that does not correspond to any realistic scenario. The risk in pure image classification or too machine–oriented tasks is to first create technologies and then create a data set for which the technology works well. This should really be the other way around and technology should adapt to the tasks (Müller et al, 2002), as otherwise the performance of a system is basically defined through the creation of the database.

One machine–oriented task that has a clear user model in mind is, for example, copy detection (Law-To et al, 2007), where distorted and modified images need to be traced back to their original. This scenario simulates a person or organization searching for copyright infringements, and similar techniques are used when uploading, for example, a video on YouTube, where Google needs to determine ex-

[6] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

tremely quickly whether copyrighted material had been used. The ImageEval benchmark had an extensive task on this topic for images and TRECVid for videos. The quality of the current techniques for copy detection tasks is generally very high.

3.2.2 User Models

For general image retrieval, a very large number of applications have been proposed (Smeulders et al, 2000; Enser, 1995) and all application domains can be used to create user models. The first domains used as user models for image retrieval are domains with a wealth of visual data available, such as journalists (Markkula and Sormunen, 1998) and librarians (Clough et al, 2005).

In terms of the application of these user models for visual information retrieval benchmarks, TRECVid first used journalists (Smeaton et al, 2003). ImageCLEF on the other hand started on the photographic retrieval task with librarians searching for images (Clough et al, 2005), then used the general public having personal photo collections (Grubinger et al, 2008), before using journalists in 2010 (Lestari Paramita et al, 2010). The choice of user model basically corresponded to the databases used. For the Wikipedia topics, general Web users of Wikipedia were taken as the user model (Tsikrika and Kludas, 2009). By having the users create the topics, while there can be influence from the researchers based on the knowledge of their own techniques, the topics created should still correspond relatively well to the user model.

ImageCLEFmed always had clinicians in mind, first with an image example, then with a clear information need regarding single images (Müller et al, 2008), and later with a specific clinical task, where similar cases were searched for (Müller et al, 2009).

For all these user models, axes can be found along which topics can be created, and along which many of the information needs can be classified. For personal photo collections, the following axes have been identified for the retrieval (Grubinger and Clough, 2007):

- temporal constraints of the retrieval, so for example during a certain period or in a certain year;
- geographical constraints such as particular places or countries;
- actions defined by the use of verbs in the queries;
- search for particular objects or persons with general nouns and proper names;
- search with adjectives that specify a characteristic of a place, object or person.

In a similar way, the following axes were found for visual information needs in the medical field:

- anatomic region (i.e. lung, liver, leg);
- imaging modality (i.e. x–ray, CT, MRI);
- pathology (i.e. fracture, cancer);
- abnormal observation (i.e. enlarged heart).

Usually much of the topic development was along these axes and normally it was checked that the information needs were not too broad and that they covered at least two of these axes.

3.2.3 Information Sources for Topic Creation

To obtain knowledge for a particular user model it is important to have access to data that underlie such information needs. In the following subsections such information sources are explained that allow for creating realistic topics, though these are mainly textual resources. This means that there is a problem in finding visual examples for realistic search topics for these user models, mainly linked to the fact that very few visual retrieval systems are in routine use. This means that the example images for the topics have to be found from other sources in addition to the textual formulation of such a user need. Such examples should of course not be part of the collection itself as otherwise the corresponding descriptions can easily be used for query expansion with a potential bias of the results

Another problem in the topic generation process is to ensure that there are relevant images in the collection for the information need. Even when the information sources for generating topics were taken into account based on the collection used, the request can still be outside of the actual content of the databases. It is thus important to develop candidate topics first, and then restrict the benchmark to a subset of these candidate topics where a sufficiently high number of relevant images can be found in the collection. The exact number of relevant images or documents is most often not important but at least a few should be findable with example search systems.

3.2.3.1 Classification Tasks

For most classification tasks within ImageCLEF such as the medical image classification task (Deselaers et al, 2007), the photo annotation task (Nowak and Dunker, 2009) and the robot vision task (Caputo et al, 2010) no dedicated topic creation is necessary as the knowledge and the type of topics are contained within the databases or the annotations of the databases. Databases are divided into training and test data and the test data are basically the topics. The exact annotation process of the databases is outside of the scope of this chapter.

These topics can still be based on user models and in the context of ImageCLEF they most often are. For the medical classification task, the user model is clinicians and the situation is that many images have either no annotation or in the case of DICOM files, the annotations are not very detailed and contain errors (Güld et al, 2002). Thus, the collection was annotated by clinicians and new images have to be annotated automatically with a chosen annotation schema based on the the training data. For the photo annotation task, several schemes were tested over the years. In

general, a collection of photographs had to be annotated with the objects or concepts contained in the images (concepts can be dogs, cars, outdoor images or night pictures, for example). Usually, a reasonably small number of concepts were chosen, typically in the range of 10–120, as current visual classification techniques often do not work very well when having to deal with a very large number of classes. Slightly different is the situation for the robot vision task, where the goal is to develop robots who can detect their own location based on the pictures they take, using training data from the same locations but under different lighting conditions and potentially with changes in the rooms such as moved furniture or modified objects. The ground truth is the location of the robot that is known and stored when recording the images.

3.2.3.2 Inherent Knowledge

The easiest way of generating topics is often to have a domain expert generate topics that correspond to a particular domain, that are challenging and at the same time useful for the chosen user model. In ImageCLEF, such an approach was taken for the first medical retrieval task (Clough et al, 2005), where a clinician very familiar with the chosen document collection selected a set of relevant images as query topics. This assured that the topics were useful and covered the collection well. On the other hand they represented the view of a single clinician and were thus not representative in this respect.

For the Wikipedia task, the inherent knowledge of the participating research groups was used (Tsikrika and Kludas, 2009), as all participants were asked to provide example topics and the topics for the evaluation were chosen from among this pool. This has an inherent risk that researchers develop topics that work well for their own system, but this risk does not bias results if all participants take part in the process. On the other hand, topics can be based too much on the technical possibilities and not on a real application of a Wikipedia user who searches for images.

3.2.3.3 Surveys and Interviews

Surveys among user groups are an important way to find out how images are being used and how visual image retrieval can help in the information retrieval process. One of the earlier studies analyzing the behavior of journalists in searching for images is described in (Markkula and Sormunen, 1998).

Within ImageCLEF, only the medical retrieval tasks used such surveys to create topics. To create the topics for the 2005 task, two surveys were performed among several groups of medical professionals in Portland (Oregon), USA and Geneva, Switzerland, (Hersh et al, 2005; Müller et al, 2006), located in medical teaching hospitals. The results of the surveys and the examples given by the experts were both used for the topic generation. The surveys also allowed definition of the differences in tasks depending on the roles of the health professionals (teaching, research, clinical work). In 2010 (Radhouani et al, 2009), another survey was performed in

Portland, OR, USA for the topic generation of ImageCLEF 2010. This time, the clinicians had access to a visual and textual retrieval system for executing example queries and analyzing the results during the interview, which can potentially give much more interesting topics and also provide image examples for the query formulation.

3.2.3.4 Usage Log Files

Log files are clearly the resource most often used as a basis for generating topics. The advantage is that they are usually available without requiring additional work and topics can thus be created just by cleaning the text in the logs. A problem with logs, particularly when they are on Web search engines, is the fact that they contain usually extremely short queries of often only one to two words, and creating a well–defined search topic from one or two terms is often hard. Library logs, such as the one used in (Clough et al, 2005) have the advantage that they contain not just a few quick terms formulated for a Web search engine, but rather well–thought–out information needs. They can thus be used more directly than Web search logs containing fewer terms. One solution to this is to add terms to make search requests more specific, or to reformulate them to reduce ambiguity and also potentially the number of relevant images. In specialized domains such as the medical field, log file terms can also be very specific with only a few search terms.

The frequency of the same search request is often used as a criterion for selection, as the most representative information needs should be used for evaluation if possible, or frequent terms should at least have a higher probability of being selected.

Concrete examples of log file use within ImageCLEF are the use of library log files of the St. Andrews library (Clough et al, 2005) for the photographic retrieval task. Other log files used for the photographic task are the Web logs of the Viventura travel agency (Grubinger et al, 2008), where the search requests were only slightly modified to be more specific and thus limit the number of relevant images. Also in the photographic task, the logs of the Belga news agency were used for topic development (Lestari Paramita et al, 2010). In all these cases, the logs corresponded to the database that was used for the retrieval.

For the medical tasks, no log files were available that correspond to the collection used for retrieval. Other information sources thus had to be found. With the health on the net media search engine[7] such a source exists and was used for ImageCLEFmed in 2006 (Müller et al, 2007). In general, some cleaning of the topics was necessary to make them more specific as most search requests were extremely general, e.g. ‘heart’ or ‘lung’. For 2007 a log file of the PubMed[8] literature search engine was used (Müller et al, 2008). This makes the selecting process more difficult as queries with visual information needs had to be found. All imaging modalities were used to pre–filter the search request and only the remaining search requests that included

[7] `http://www.hon.ch/HONmedia/`

[8] `http://www.pubmed.gov/`

Table 3.1: Sources used for generating the query topics in ImageCLEF (not including the interactive and geographic query tasks).

Year	2003	2004	2005	2006	2007	2008	2009
Photo retrieval	St. Andrews logs	St. Andrews logs	St. Andrews logs	viventura web logs	viventura web logs	viventura web logs	Belga logs
Photo Annot.			annotated data	annotated data	annotated data	annotated data	annotated data
Medical retrieval		expert knowledge	expert survey	web logfile HON	Medline queries	from previous years	expert survey
Medical Annot.			annotated data	annotated data	annotated data	annotated data	annotated data
Nodule detection							expert annotations
Wikipedia						user generated	user generated
Robot vision						places known	places known

a modality were taken into consideration for the topic development based on the frequency of their occurrence.

3.3 Concrete Examples for Generated Visual Topics in Several Domains

This chapter gives a few examples for topics created in the context of ImageCLEF tracks using the various sources described. Table 3.1 also gives an overview of the ImageCLEF tasks and their way of generating the topics over the seven years of ImageCLEF. It can be seen that all purely visual tasks used only annotated data for generating topics and relevance judgments. This means that the tasks are really classification and not retrieval tasks, and the separation of the data into test data and training data was usually done in a more or less random fashion that took into account a certain distribution among training and test data.

By contrast the Wikipedia task used participant–generated topics, and the photographic retrieval task used three different types of log files. The medical retrieval task changed the topic generation almost every year using first expert knowledge, then user surveys and then two different types of log files for the topic generation. It is not possible to give examples for all tasks in this chapter and the corresponding Chapters 7, 8, 9, 10, 11, 12, and 13 can be used to find further details about each of the tasks.

3.3.1 Photographic Retrieval

In the Wikipedia task the topics were generated by the participants of the task as described by Tsikrika and Kludas (2009). In Figure 3.1, an example for such a topic

```
<topic>
<number> 1 </number>
<title> cities by night <title>
<image> hksky2.jpg </image>
<narrative> I am decorating my flat and as I like photos
  of cities at night, I would like to find some that I could
  possibly print into posters. I would like to find photos of
  skylines or photos that contain parts of a city at night
  (including streets and buildings). Photos of cities
  (or the earth) from space are not relevant.
</narrative>
</topic>
```

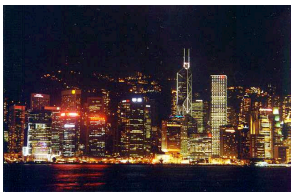

Fig. 3.1: Example topic for the Wikipedia task including a visual example, a title and a narrative describing the detailed information need.

can be seen. For the retrieval, the participating research groups could decide to use only the title, or to include the narrative as well. An image was supplied for almost all topics in the first year as can be seen in Figure 3.1, whereas in subsequent years several images were supplied for each topic.

The practice of using task participants for generating the topics was taken from the INEX[9] multimedia track (Westerveld and van Zwol, 2007) and has worked well over the years.

For the ImageCLEF photo retrieval retrieval task, various log files have been used over the years for generating the topics. An example for a topic using the Viventura log file can be seen in Figure 3.2. Several example images were supplied with each of the topics. In addition to the title and the narrative, the language of the topics can vary between German, English and Spanish. The user model is a person having a large personal collection of holiday pictures.

3.3.2 Medical Retrieval

An overview for medical image retrieval and its applications is given by Müller et al (2004). The topic developments for ImageCLEFmed generally modeled a clinician working on a particular case and who had a specific information need. Other roles of clinicians such as teacher and researcher were also considered. Figure 3.3 shows

[9] INitiative for the Evaluation of XML retrieval, `http://www.inex.otago.ac.nz/`

```
<top>
<num> Number: 14 </num>
<title> scenes of footballers in action </title>
<narr> Relevant images will show football (soccer)
players in a game situation during a match. Images with
footballers that are not playing (e.g. players posing for
a group photo, warming up before the game, celebrating
after a game, sitting on the bench, and during the half-
time break) are not relevant. Images with people not
playing football (soccer) but a different code (American
Football, Australian Football, Rugby Union, Rugby League,
Gaelic Football, Canadian Football, International Rules
Football, etc.) or some other sport are not relevant.
</narr>
<image> images/31/31609.jpg </image>
<image> images/31/31673.jpg </image>
<image> images/32/32467.jpg </image>
</top>
```

Fig. 3.2: Examples topic from the photographic retrieval task.

an example topic. Topics were always supplied in three languages (English, French, German) and with several example images. Topics were also developed along the axes anatomy, modality, pathology and abnormality. In the case of the topic shown, the two axes modality (x–ray) and pathology (fractures) are covered.

Due to the large variety of potential results of all anatomic regions in this case, the query can not be considered a visual query as it cannot be solved with visual features alone. It is thus regarded as a mixed query as visual features can help to distinguish x–ray images from other modalities.

3.4 The Influence of Topics on the Results of Evaluation

The various examples and ways of creating topics have shown that topic development is not an easy process. This raises the question of why invest a large amount of time and effort into creating such topics? The answer is that the entire evaluation that follows in an evaluation campaign or a single system evaluation is based on the topics developed. The topics have a much stronger influence on the comparative evaluation than the database itself and the relevance judgments have. Thus,

```
Show me all x-ray images showing fractures.
Zeige mir Roentgenbilder mit Bruechen.
Montres-moi des radiographies avec des fractures.
```

Fig. 3.3: A query requiring more than visual retrieval but where visual features can deliver hints to good results.

the importance of the topic development should not be taken lightly and it needs to be made clear what the main goal in the topic development is. It has to be clearly stated whether the topic development is based on any real application, or whether the capabilities of a certain technique are to be tested mainly in a laboratory style evaluation. Very often topics pretend to be modeling real–world applications when they are really not doing so.

3.4.1 Classifying Topics Into Categories

To further analyze information retrieval techniques, the topics can be classified into groups that can subsequently be used for analyzing techniques separately. Within several ImageCLEF tasks, the topics are classified into visual, textual and mixed topics by an experienced researcher in the field. This allows us to separately measure the best techniques for each of these categories.

Grubinger and Clough (2007) surveyed several of the ImageCLEFphoto topics for their level of 'visualness' (very bad, bad, average, good, very good). Several researchers judged the topics with respect to the visualness and then compared the performance results using a visual system for retrieval, showing that visualness can be estimated very well.

Topics can also be classified into other categories, allowing us to separately analyze the influences of certain techniques for particular tasks (e.g. tasks with a geographical orientation, topics with actions, topics of particular persons or topics with temporal constraints).

3.4.2 Links Between Topics and the Relevance Judgments

As the concept of relevance in retrieval tasks is not very stable, there are several approaches for linking the topic creation process with the relevance judgement process. In TREC, the people creating the topics are usually the people who also judge the pools for relevance. This has the advantage that the topic creator knows what he had in mind with the task creation, but on the other hand this can be very different if another person is judging the same topic. In the Wikipedia task, part of the topic creation and relevance judgement process is also performed by the participants and thus potentially by the topic creators. In the medical tasks of ImageCLEF, domain experts judge the topics but have not created the topics themselves. In general, several people judge the same topics, which allows us to analyze the level of ambiguity in the topic. This also allows us to find out whether the topic was well formulated for the system, and potentially ambiguous topics can still be removed at this point.

An extremely important step when developing topics with judgment in mind is to have a very detailed description or narrative of the task. Particularly if the relevance judges have not created the topics themselves it is important to detail exactly what is to be regarded as relevant. A description of exactly what is regarded as non–relevant is also extremely important as this can help define the border between relevant and non–relevant documents or images. The descriptions for the relevance judgements of the medical task have grown to over five pages, meaning they detail the entire process and define where the border between relevant and non–relevant is.

3.4.3 What Can Be Evaluated and What Can Not?

One of the questions is also with respect to what the limit of system capabilities is that can be evaluated. Jörgensen (1999) details the limits of image retrieval systems with respect to emotions, feelings and impressions but also shows ways how this can at least partially be reached. It is clear that query topics in image retrieval benchmarks need to correspond to current system capabilities and need to propose challenging search problems for the research community. To continue proposing challenging problems it is extremely important to have the topics evolve regularly over time, for example making them more challenging. If the topics of the benchmarks do not evolve sufficiently, the participating teams can be over–optimized for a particular scenario and this has to be avoided. The photo retrieval task has in this context evolved in several directions from evaluating very large databases to evaluating diversity. For the medical task this has been the creation of much larger data sets and also the development from image retrieval to case–based retrieval including images. This evolution has to be retained although it usually means additional work for the participants and also reduces the number of research groups participating as participation means increased work.

Another concept that can be important for generating topics is the concept of diversity. This was used in ImageCLEF for the photographic retrieval task in 2008

and 2009 (Lestari Paramita et al, 2010). In this case not only the topics need to be created but also the clusters of images for each topic that correspond to different representations of a particular search topic.

3.5 Conclusions

Topic creation is an important part of the evaluation of information retrieval systems, especially for visual information retrieval. As systems start to reach a quality where they can be used in real applications, mainly when used in combination with text retrieval, it is important to prove the quality of the tools. For this it is important to direct research efforts towards real problems and scenarios where image retrieval can deliver an added value. For this it seems necessary to have clear user models in mind, then create databases and topics based on the user models and then optimize techniques for these topics and databases. This avoids optimizing the data set to deliver good results for a particular technique (Müller et al, 2002), and so advances the technology.

Topic development is important for the creation of information retrieval tasks and more effort is necessary to control all the variables in this process. Parameters such as topic difficulty, topic variety and particularly the orientation towards real problems has to be taken into account to advance image retrieval through using good evaluation practices.

In the context of cross–language information retrieval it also needs to be stated that image retrieval offers a valuable contribution to language–independent information retrieval, as annotations with concepts can generate annotations in any language. Visual image analysis can also find similar images independent of the language. Within ImageCLEF several tasks are totally language–independent whereas others use collections in English and then propose topics in several languages. Starting from 2010 Wikipedia will have images annotated in various languages, which is the norm in the context of Wikipedia where content is created in many languages. Such a scenario can actually increase the importance of visual retrieval that currently has poorer performance than textual image retrieval.

Acknowledgements I would like to thank all funding institutions who have made this work possible, notably the European Union through the 6th and 7th framework program through the MultiMatch, TrebleCLEF, Chorus+, Promise and Khresmoi projects, the American National Science Foundation (NSF), the Swiss National Science Foundation (SNF), Google, the University of Geneva and the University of Applied Sciences Western Switzerland (HES–SO). Thanks are also due to all those who made their data available to ImageCLEF.

References

Caputo B, Pronobis A, Jensfelt P (2010) Overview of the CLEF 2009 robot vision task. In: Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers, Springer, Lecture Notes in Computer Science (LNCS)

Cleverdon CW (1962) Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Tech. rep., Aslib Cranfield Research Project, Cranfield, USA

Clough PD, Müller H, Sanderson M (2005) The CLEF cross–language image retrieval track (ImageCLEF) 2004. In: Peters C, Clough PD, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign, Springer, Bath, UK, Lecture Notes in Computer Science (LNCS), vol 3491, pp 597–613

Clough PD, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 cross–language image retrieval track. In: Cross–Language Evaluation Forum (CLEF 2005), Springer, Lecture Notes in Computer Science (LNCS), pp 535–557

Deselaers T, Müller H, Clough PD, Ney H, Lehmann TM (2007) The CLEF 2005 automatic medical image annotation task. International Journal in Computer Vision 74(1):51–58

Enser PGB (1995) Pictorial information retrieval. Journal of Documentation 51(2):126–170

Everingham M, Zisserman A, Williams CKI, van Gool L, Allan M, Bishop CM, Chapelle O, Dalal N, Deselaers T, Dorko G, Duffner S, Eichhorn J, Farquhar JDR, Fritz M, Garcia C, Griffiths T, Jurie F, Keysers D, Koskela M, Laaksonen J, Larlus D, Leibe B, Meng H, Ney H, Schiele B, Schmid C, Seemann E, Shawe-Taylor J, Storkey A, Szedmak S, Triggs B, Ulusoy I, Viitaniemi V, Zhang J (2006) The 2005 PASCAL visual object classes challenge. In: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment (PASCAL Workshop 05), Southampton, UK, no. 3944 in Lecture Notes in Artificial Intelligence (LNAI), pp 117–176

Goodrum A (2000) Image information retrieval: An overview of current research. Informing Science 3(2):63–66

Grubinger M, Clough PD (2007) On the creation of query topics for ImageCLEFphoto. In: MUSCLE/ImageCLEF workshop 2007, Budapest, Hungary

Grubinger M, Clough P, Hanbury A, Müller H (2008) Overview of the ImageCLEF 2007 photographic retrieval task. In: CLEF 2007 Proceedings, Springer, Budapest, Hungary, Lecture Notes in Computer Science (LNCS), vol 5152, pp 433–444

Güld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM (2002) Quality of DICOM header information for image categorization. In: International Symposium on Medical Imaging, San Diego, CA, USA, SPIE Procweedings, vol 4685, pp 280–287

Gunther NJ, Beretta G (2001) A benchmark for image retrieval using distributed systems over the Internet: BIRDS–I. Tech. rep., HP Labs, Palo Alto, Technical Report HPL–2000–162, San Jose

Harman D (1992) Overview of the first Text REtrieval Conference (TREC–1). In: Proceedings of the first Text REtrieval Conference (TREC–1), Washington DC, USA, pp 1–20

Hersh W, Jensen J, Müller H, Gorman P, Ruch P (2005) A qualitative task analysis for developing an image retrieval test collection. In: ImageCLEF/MUSCLE workshop on image retrieval evaluation, Vienna, Austria, pp 11–16

Jörgensen C (1999) Retrieving the unretrievable in electronic imaging systems: emotions, themes and stories. In: Rogowitz B, Pappas TN (eds) Human Vision and Electronic Imaging IV, San Jose, California, USA, SPIEProc, vol 3644, (SPIE Photonics West Conference)

Law-To J, Joly LCA, Laptev I, Buisson O, andNozha Boujemaa VGB, Stentifordl F (2007) Video copy detection: a comparative study. In: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM press, pp 371–378

Lestari Paramita M, Sanderson M, Clough P (2010) Diversity in Photo Retrieval: Overview of the ImageCLEFphoto Task 2009. In: Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones JFG, Gonzalo J, Caputo B (eds) Multilingual Information Access Evaluation Vol. II Multimedia

Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers, Corfu, Greece, Lecture Notes in Computer Science (LNCS)

Markkula M, Sormunen E (1998) Searching for photos — journalists' practices in pictorial IR. In: Eakins JP, Harper DJ, Jose J (eds) The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval, The British Computer Society, Newcastle upon Tyne, Electronic Workshops in Computing

Mizzaro S (1997) Relevance: The whole (hi)story. Journal of the American Society for Information Science 48(9):810–832

Müller H, Müller W, Squire DM, Marchand-Maillet S, Pun T (2001) Performance evaluation in content–based image retrieval: Overview and proposals. PRL 22(5):593–601, special Issue on Image and Video Indexing

Müller H, Marchand-Maillet S, Pun T (2002) The truth about Corel – Evaluation in image retrieval. In: Lew MS, Sebe N, Eakins JP (eds) Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002), Springer, London, England, Lecture Notes in Computer Science (LNCS), vol 2383, pp 38–49

Müller H, Michoux N, Bandon D, Geissbuhler A (2004) A review of content–based image retrieval systems in medicine—clinical benefits and future directions. International Journal of Medical Informatics 73(1):1–23

Müller H, Despont-Gros C, Hersh W, Jensen J, Lovis C, Geissbuhler A (2006) Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006), Maastricht, The Netherlands, IOS Press, Studies in Health Technology and Informatics, pp 24–32

Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A (2007) Analyzing web log files of the Health On the Net HONmedia search engine to define typical image search tasks for image retrieval evaluation. In: MedInfo 2007, Brisbane, Australia, IOS press, Studies in Health Technology and Informatics, vol 12, pp 1319–1323

Müller H, Deselaers T, Kim E, Kalpathy-Cramer J, Deserno TM, Clough PD, Hersh W (2008) Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings, Springer, Budapest, Hungary, Lecture Notes in Computer Science (LNCS), vol 5152, pp 473–491

Müller H, Kalpathy-Cramer J, Hersh W, Geissbuhler A (2008) Using Medline queries to generate image retrieval tasks for benchmarking. In: Medical Informatics Europe (MIE2008), IOS press, Gothenburg, Sweden, pp 523–528

Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Said R, Bakke B, Kahn Jr CE, Hersh W (2009) Overview of the CLEF 2009 medical image retrieval track. In: Working Notes of CLEF 2009, Corfu, Greece

Nowak S, Dunker P (2009) Overview of the CLEF 2009 Large–Scale Visual Concept Detection and Annotation Task. In: Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones J, Gonzalo J, Caputo B (eds) Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers, Corfu, Greece, LNCS

Radhouani S, Kalpathy-Cramer J, Bedrick S, Hersh W (2009) Medical image retrieval, a user study. Tech. rep., Medical Inforamtics and Outcome Research, OHSU, Portland, OR, USA

Savoy J (2002) Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum), Springer, Darmstadt, Germany, Lecture Notes in Computer Science (LNCS), vol 2406, pp 27–43

Smeaton AF, Kraaij W, Over P (2003) TRECVID 2003: An overview. In: Proceedings of the TRECVID 2003 conference

Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12):1349–1380

Sparck Jones K, van Rijsbergen C (1975) Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge

Tsikrika T, Kludas J (2009) Overview of the wikipediaMM task at ImageCLEF 2008. In: Peters C, Giampiccolo D, Ferro N, Petras V, Gonzalo J, Peñas A, Deselaers T, Mandl T, Jones G, Kurimo M (eds) Evaluating Systems for Multilingual and Multimodal Information Access — 9th Workshop of the Cross–Language Evaluation Forum, Aarhus, Denmark, Lecture Notes in Computer Science (LNCS)

Voorhees EM, Harmann D (2000) Overview of the ninth Text REtrieval Conference (TREC–9). In: The Ninth Text Retrieval Conference, Gaithersburg, MD, USA, pp 1–13

Westerveld T, van Zwol R (2007) The INEX 2006 Multimedia track. In: Fuhr N, Lalmas M, Trotman A (eds) Comparative Evaluation of XML Information Retrieval Systems, Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006), Revised Selected Papers, Springer, Lecture Notes in Computer Science (LNCS), vol 4518, pp 331–344