Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012

David Zellhöfer

Brandenburg Technical University, Database and Information Systems Group,
Walther-Pauer-Str. 1, 03046 Cottbus
david.zellhoefer@tu-cottbus.de

Abstract. As a consequence of a discussion at ImageCLEF 2011, the personal photo retrieval pilot task has been designed to represent a personal photo collection. In contrast to other existing collections where the contributors often remain unknown, the proposed collection has been sampled from 19 layperson photographers and enriched by their demographics.

To ensure a variance in photographic motifs and style, the contributors have been chosen from different demographic groups. Thus, one can interpret the content of the collection as a mirror of a photographer's lifespan with typical changing usage behaviors, cameras, topics, and places.

The task consists of two subtasks. The first task is aiming at retrieving visual concepts such as trees, animals, or market scenes. The second is focussing on the retrieval of particular events such as parties or rock concerts. To solve both tasks, the participants were provided with query-by-example documents in addition to browsing data.

The participation in this task was very low as only three groups submitted results. To summarize the first subtask, the best group achieved a precision at 20 of 0.7333 and a NDCG at 20 of 0.5459. In contrast, the second subtask focussing on events was solved with a precision at 20 of 0.9333 and a NDCG at 20 of 0.9697.

Regarding the provided browsing data, only one group decided to exploit this resource instead of the provided metadata. Interestingly, it could use this data successfully to solve subtask 1 but reached the last position at subtask 2. This result indicates that there is a particularly strong influence of metadata on the retrieval of events.

Keywords: Content-Based Image Retrieval, Benchmark, Experiments, Personal Photograph Collection

1 Introduction

As a consequence of a discussion at ImageCLEF 2011, the personal photo retrieval pilot task has been designed to represent a personal photo collection. The presented pilot task is aiming at providing a test bed for QBE-based retrieval scenarios in the scope of personal information retrieval. In contrast to other tasks relying on downloads from Flickr or the like, the underlying data set reflects an amalgamated personal image collection that has been taken by 19 photographers. Hence, it can be used best as a test set for layperson retrieval tasks carried out ad hoc on their own collections such as: “find all images with a street scene”, “find a beach similar to this”, or more event-based tasks like “show me more pictures from the last U2 concert”. The aim of this pilot task is to retrieve relevant images based on typical layperson usage scenarios in their own collections, i.e., the search for similar images or images depicting a similar event, e.g. a rock concert [6].

To ensure a variance in photographic motifs and style, the contributors have been chosen from different demographic groups. Thus, one can interpret the content of the collection as a mirror of a photographer’s lifespan with typical changing usage behaviors, cameras, topics, and places.

Unlike system-centric (Cranfield-based) benchmarks, the pilot tasks tries to establish a more user-centered perspective on multimodal information retrieval (MIR) and content-based image retrieval (CBIR). As such, it features two different retrieval subtasks that can be derived from the camera usage behavior of the contributing photographers (see below). Additionally, it provides simulated browsing data reflecting a user’s interaction with the system based on multiple search strategies as observed by [4] or described by [1] respectively.

In order to express the subjectivity of relevance assessments, the ground truth is based on graded relevance judgements. To include these assessments into the evaluation, the pilot tasks uses the NDCG metric [3] (see Section 4.1) in addition to precision at various cut-off levels.

As said before, the task consists of two subtasks. The first task is aiming at retrieving visual concepts such as trees, animals, or market scenes. The second is focussing on the retrieval of particular events such as parties or rock concerts. To solve both tasks, the participants were provided with query-by-example documents in addition to browsing data.

2 Task Resources

The pilot task relies on a subset of the Pythia dataset [6] which will be described in the next section. To complete the description of the provided resources, Section 2.2 will comment on the acquisition of the ground truth. The following section will then discuss the elicitation of the browsing data offered to the participants as an additional resource.

2.1 The Pythia Dataset

To overcome limitations by binary relevance judgments often found in common test collections, the Pythia collection [6] has been proposed. The collection is aiming at providing a benchmark for user-centered or relevance feedback-related experiments which are affected by subjective relevance levels in particular. The collection differs from collections consisting of Flickr downloads or the like as it has been sampled from 19 layperson photographers. For the individual contribution of the photographers, see Figure 1. In addition to the image data, the contributors to the collection completed a survey asking for their photograph taking behavior, their demographics etc. To ensure a variance in photographic motifs and style, the contributors have been chosen from different demographic groups. Thus, one can interpret the content of the collection as a mirror of a photographer's lifespan with typical changing usage behaviors, cameras, topics, and places. The total size of the collection is 5,555 documents.

The documents within the collection have neither been processed extensively nor have duplicates been removed. Hence, the data can be considered a realistic sample from a typical user's hard-disk. The collection is rich on metadata including GPS, IPTC, EXIF, and information about events depicted on each photography. All this information is available to the participants of the pilot task. For an overview, see Table 1.

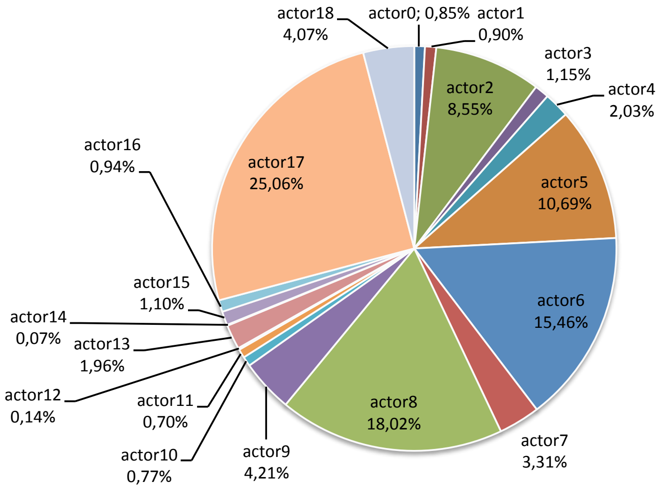


Fig. 1. Photographers' Contribution in Percent [6]

Table 1. Metadata Characteristics (Excerpt) [6]

Characteristic	%
EXIF (Date, Camera Info. etc.)	100.00
GPS Data	81.85
Event Tags	96.71
Outdoor Photographies	82.64
Indoor Photographies	17.41

2.2 Ground Truth Acquisition

In order to obtain the ground truth, 42 assessors were asked to participate. The core characteristics can be subsumed as follows. The majority of the assessors (28 out of 42) are male and born between 1979 and 1991 (median: 1987). Most of the assessors are students with a background in economics (26), the second largest group (13) has a background in computer science and information technology. Figure 2 illustrates the other fields of education or working area. Regarding their level of expertise in the field of MIR or IR, 9 assessors took classes in MIR while 11 heard IR. When asked directly about their knowledge of the field the median lies at “little knowledge” with an average of 1.40, i.e., a trend towards considering themselves as an ‘informed outsiders”.

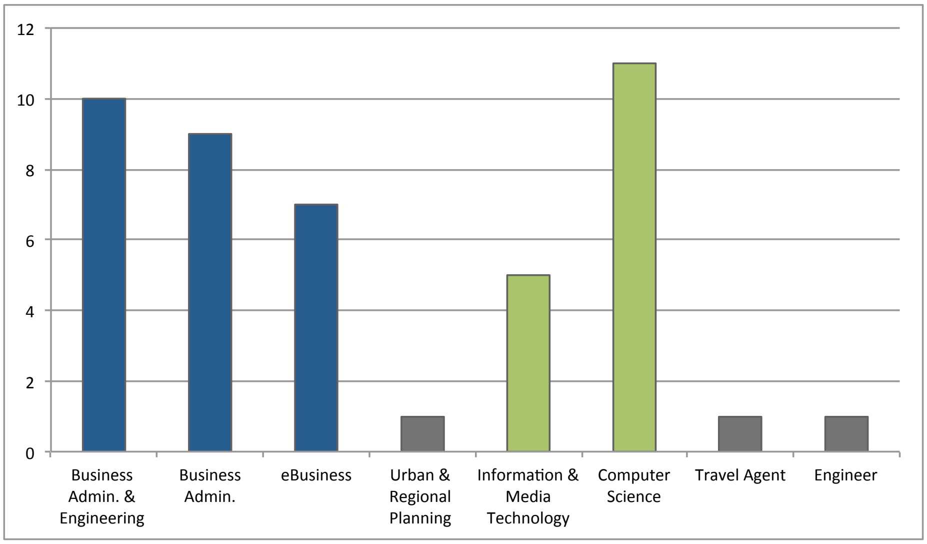
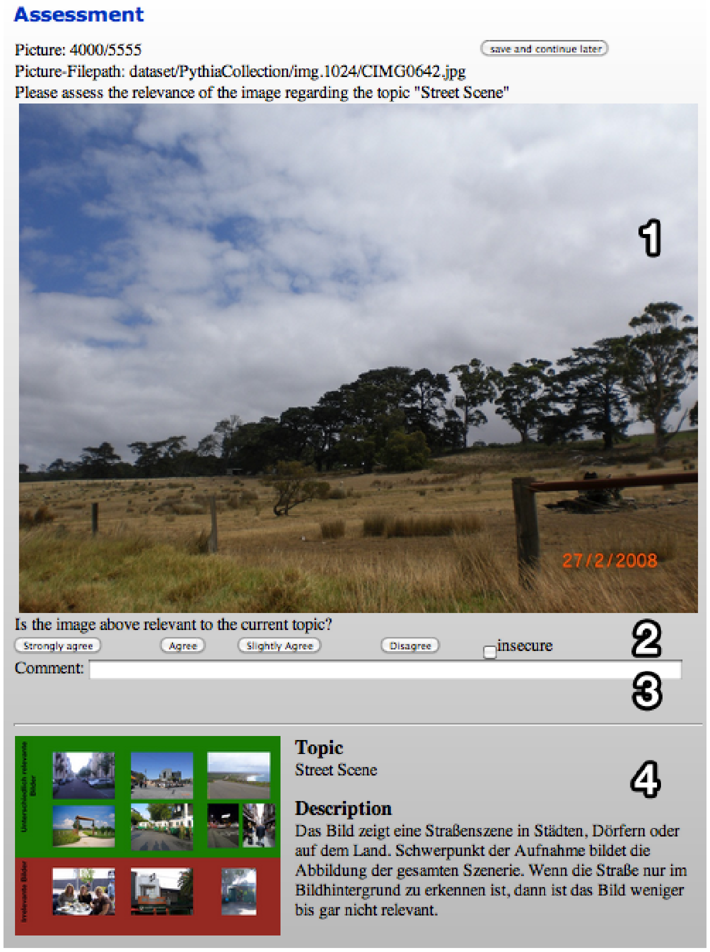


Fig. 2. Job types of the assessors; economic background is marked blue, IT is green

Using a web-based evaluation tool (see Figure 3), the assessors could judge the relevance of an image with respect to a topic on a graded scale ranging from 0 (irrelevant) to 3 (fully relevant). All assessors had to judge all documents regarding a topic. The topics were associated with the assessors by random. To keep them motivated, the assessors were allowed to work with the collection from a place of their choice. Additionally, they could pause an assessment run and continue from later on. A time constraint has not been defined. In average 2.69 topics were evaluated per assessor (standard deviation: 1.60). The individual assessments were saved separately in order to maintain them for later usage.

Calculation of the Ground Truth for each Topic Based on the individual assessments, an averaged ground truth has been calculated. First, the frequency of each graded relevance judgement (out of an interval from 0 (irrelevant) to 3 (fully relevant)) was counted per image and topic. Based on these relevance judgment frequencies, an estimation value was calculated and rounded. The rounded estimation value of the relevance of an image regarding a topic was then used as the averaged graded relevance assessment for this image. In consequence, each image could be associated with a graded relevance judgment for each topic.



Assessment

Picture: 4000/5555

Picture-Filepath: dataset/PythiaCollection/img.1024/CIMG0642.jpg

Please assess the relevance of the image regarding the topic "Street Scene"

Is the image above relevant to the current topic?

Comment:

Topic

Street Scene

Description

Das Bild zeigt eine Straßenszene in Städten, Dörfern oder auf dem Land. Schwerpunkt der Aufnahme bildet die Abbildung der gesamten Szenerie. Wenn die Straße nur im Bildhintergrund zu erkennen ist, dann ist das Bild weniger bis gar nicht relevant.

Fig. 3. Web-based assessor's GUI; 1) current sample image, 2) graded relevance scale, 3) free text comment field, 4) task description with relevant and irrelevant images

Generating Browsing Information As we could not obtain real browsing information, it had to be generated artificially. Using the graded relevance assessments, multiple images were chosen as browsing images. The provided browsed images have a relevance grade ranging from 1 to 2, i.e., they are judged neither irrelevant nor fully relevant for a given topic. In other words, the browsing data consists of interesting images which were not fully relevant for the modeled user which caused him or her to proceed with the search. This change of search strategy (from browsing to a directed QBE search) is reflected by the following subtask.

3 Task Description

3.1 Subtask 1: Retrieval of Visual Concepts

The objective of the first subtask is to find similar images to a specified visual concept or topic. Out of the 32 topics provided by the Pythia dataset [6], the 24 topics with the most relevant images in the corpus were chosen. The topics are listed in Table 2.

To solve the task, 5 QBE documents were provided to the participants. All QBE documents are fully relevant according to our assessors. In addition to the metadata present in the images, browsing data consisting of images that have been inspected during the search (see above) is offered. The usage of this browsing data is voluntary as the utilization of image features or metadata (e.g. GPS information).

Table 2. Topics of subtask 1; bold set topics indicate used topics

1.	**Beach and Seaside**	17.	Still Life
2.	**Street Scene**	18.	**Church (Christian)**
3.	**Statue and Figurine**	19.	**Art Object**
4.	**Asian Temple** & Palace	20.	**Cars**
5.	**Landscape**	21.	**Ship / Maritime Vessel**
6.	Hotel Room	22.	Airplane
7.	People	23.	**Temple (Ancient)**
8.	**Architecture (profane)**	24.	**Squirrels**
9.	**Animals**	25.	**Sign**
10.	**Asian Temple Interior**	26.	**Mountains**
11.	**Flower / Botanic Details**	27.	Monkeys
12.	Market Scene	28.	**Birds**
13.	**Submarine Scene**	29.	**Trees**
14.	**Ceremony and Party**	30.	Abstract Content
15.	**Theater / Performing Arts**	31.	**City Panorama**
16.	**Clouds**	32.	(Christian) Church Interior

3.2 Subtask 2: Retrieval of Events

With respect to the fact that most contributors to the collection used their cameras only at special events [6], an additional event retrieval subtask was defined. Its objective is to find further images from an event specified by 3 QBE images from the same event. In contrast to subtask 1, browsing data is not available. Table 3 lists all events.

The events range from special events such as a U2 concert to their generalization, i.e., a rock concert. It is noteworthy that the events can reoccur and are not always chronologically connected. The focus on events representing a holiday or a city trip is not a freely chosen bias. Instead, it reflects the state of randomly picked images from real-world personal photo collections [6].

Table 3. Topics of subtask 2

1.	Conference	9.	Hamburg Holiday
2.	Fire	10.	London Holiday
3.	Excursion	11.	Party
4.	Flight	12.	U2 Concert
5.	Australia Holiday	13.	Scuba Diving
6.	Bali Holiday	14.	Rock Concert
7.	Egypt Holiday	15.	Mountainside Holiday
8.	Greece Holiday		

4 Results

4.1 Evaluation Metrics

It is widely known that relevance judgments are highly subjective. Because of this fact, the presented ground truth is based on a gradual scale of relevance. Unfortunately, traditional measurements such as the mean average precision (MAP) or precision at n cannot deal with this kind of judgements. Hence, we will rely on the discounted cumulative gain (DCG) measurement [3] in addition to precision at n. As stated in [6] "DCG relies on graded relevance assessments and has become more and more used within the information retrieval (IR) community, which is reflected by a performance evaluation of different metrics presented at SIGIR '11 showing that DCG 'really is a useful user-centered measure of system effectiveness' [2]. Besides its capability of reflecting subjectivity, DCG also provides more appropriate means to evaluate relevance feedback (RF) or adaptive systems as it can be used to measure slight changes or re-orderings of relevant documents with varying degrees of relevance within the result list". The core idea of DCG is to apply "a discount factor to the relevance scores in order to devaluate late-retrieved documents" [3]. In other words, the metric rewards highly relevant documents at the first positions in the result ranking and punishes systems retrieving less relevant documents at the first places. For the scope of this task, the DCG implementation of `trec_eval` version 9.0 with standard discount settings is used. A full discussion of the metric is available by Järvelin and Kekäläinen [3].

4.2 Results of the Participants

Because of the low participation rate, a general interpretation of the results is hardly possible. Table 4 and 5 summarize the participants' results. The submitted runs consist of both automatic and manually assisted runs. While two groups worked without relevance feedback (NOFB), the University of Cagliari used it in a binary way. That is, relevance feedback was given with relevance or irrelevance judgments.

Regarding the retrieval type, the runs are more diverse. The participants could use the following combinations of the provided data and metadata:

- visual features alone (IMG)
- visual features and metadata (IMGMET)
- visual features and browsing data (IMGBRO)
- metadata alone (MET)
- metadata and browsing data (METBRO)
- browsing data alone (BRO)
- a combination of all modalities (IMGMETBRO)

None of the participants used all modalities in combination. The participants relied on IMG, MET, IMGMET, or IMGBRO alone. Interestingly, only the group REGIM decided to exploit the browsing data instead of the provided metadata. Surprisingly, it could use this data successfully to solve subtask 1 but reached the last position at subtask 2. This result indicates that there is a particularly strong influence of metadata on the retrieval of events.

To summarize the first subtask (see Table 4), the best group achieved a precision at 20 of 0.7333 and a NDCG at 20 of 0.5459. In contrast, the second subtask focussing on events was solved with a precision at 20 of 0.9333 and a NDCG at 20 of 0.9697 (see Table 5).

The Effect of Different User Groups on the Retrieval Quality Because of the nature of the acquisition of the ground truth (see Section 2.2), distinct ground truths could be generated per user groups. The main objective for these different ground truths was to examine if the retrieval metrics for each participant differ per user group. Hence, 6 user groups were defined on basis of the demographics of the assessors. These are:

Experts A group of users that stated that they have an expertise with IR.
Non-Experts The complement of the experts group.
Male/Female The assessors divided by gender.

IT This groups consists of assessors with an IT background (see Figure 2).
Non-IT The complement of the IT group.

As not all images and topics have been assessed by members of each separate user group, missing assessments had to be added from the averaged ground truth (see above). Figures 4-6 illustrate the results of some sample runs regarding different user groups. The x-axis indicates different retrieval measurements, i.e., 1) P@10, 2) P@20, 3) P@30, 4) NDCG@10, 5) NDCG@20, and 6) NDCG@30. Besides in Figure 6, the results of the individual user groups are very close to the results from the averaged ground truth. Further research is needed to find out why this is the case. For now, it seems that the addition of missing relevance assessments is causing this low level of variation.

5 Conclusions and Future Work

As this is the pilot phase of a more user-centered benchmark, the task posed more questions and revealed more issues as it actually answered.

First, it became obvious that the generation of user-centered tasks and the acquisition of the accompanying data takes much more time than expected. Originally, we also wanted to provide data for user simulations to all participants so they could tune their systems with respect to different user groups. Due to the time constraints, this data could not be released on time. If this has had an impact on the low participation rate remains an open question.

To our surprise, only one group used the provided browsing data. Regarding this data, we expected more interest as studies in interactive IR clearly show that users are changing their search strategies during the search process [4]. Anyhow, the positive results of this group might motivate further studies of others how to exploit this resource.

Interestingly, there was no interest in solving the so-called user-centered initiative of the subtasks. The initiative asked for an alternative representation of the top-k results offering a more diverse view onto the results to the user. This challenge reflects the assumption that a user-centered system should offer users good and varying retrieval results. Varying results are likely to compensate for the vagueness inherent in both retrieval and query formulation. Hence, an additional filtering or clustering of the result list could improve the effectiveness and efficiency (in terms of usability) of the retrieval process. It remains unclear, if this task was too complex or just out of the area of expertise of the participants that used the dataset for the first time.

To conclude with, we are happy that the participants tried to solve to task using diverse techniques and hope to motivate further research in the field of user-centered MIR and CBIR.

Acknowledgments

This research was supported by a grant of the Federal Ministry of Education and Research (Grant Number 03FO3072).

Table 4. Results of subtask 1 (excerpt); bold values indicate the best result

Group	**Run ID**	**Run Type**	**Relevance Feedback**	**Retrieval Type**	**P_20**	**ndcg_cut_20**
KIDS	IBMA0	Automatic	NOFB	IMGMET	0.6896	**0.5459**
KIDS	OBOA0	Automatic	NOFB	MET	0.6354	0.4836
KIDS	IOMA0	Automatic	NOFB	IMGMET	0.6104	0.4872
KIDS	OBMA0	Automatic	NOFB	MET	0.5771	0.4066
REGIM	run4	Automatic	NOFB	IMGBRO	**0.7333**	0.4563
REGIM	run2	Automatic	NOFB	IMGBRO	0.7292	0.4561
REGIM	run1	Automatic	NOFB	IMGBRO	0.7292	0.456
REGIM	run5	Automatic	NOFB	IMGBRO	0.7292	0.4551
REGIM	run3	Automatic	NOFB	IMGBRO	0.7292	0.4551
University of Cagliari	Run_1_2	Feedback and/or human assistance	BINARY	IMG	0.6938	0.5457
KIDS	IOOA4	Automatic	NOFB	IMG	0.5354	0.4545
University of Cagliari	Run_1_1	Feedback and/or human assistance	BINARY	IMG	0.5646	0.4835
University of Cagliari	Run_3_2	Feedback and/or human assistance	BINARY	IMG	0.3958	0.3466
				Mean	0.6425	0.4640
				Std. Dev.	0.1028	0.0518

Table 5. Results of subtask 2 (excerpt); bold values indicate the best result

Group	**Run ID**	**Run Type**	**Relevance Feedback**	**Retrieval Type**	**P_20**	**ndcg_cut_20**
KIDS	OOMA0	Automatic	NOFB	MET	**0.9333**	**0.9697**
KIDS	IOMA0	Automatic	NOFB	IMGMET	0.9267	0.9655
KIDS	IOMA0-2	Automatic	NOFB	IMGMET	0.8100	0.8636
KIDS	IOMA0-3	Automatic	NOFB	IMGMET	0.7867	0.8357
KIDS	IOOA0	Automatic	NOFB	IMG	0.4833	0.5446
REGIM	run8	Automatic	NOFB	IMGBRO	0.1767	0.1936
REGIM	run7	Automatic	NOFB	IMGBRO	0.1733	0.1915
REGIM	run9	Automatic	NOFB	IMGBRO	0.1733	0.1913
				Mean	0.5579	0.5944
				Std. Dev.	0.3463	0.3580

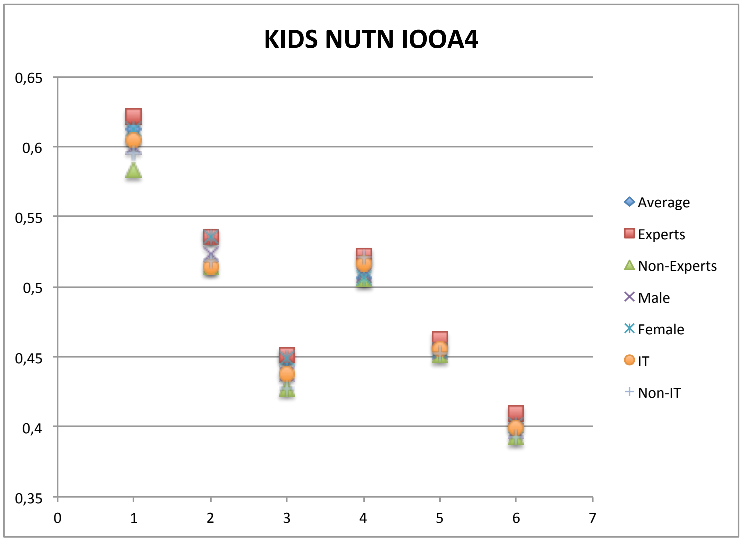


Fig. 4. Retrieval performance for different user groups (KIDS NUTN IOOA4)

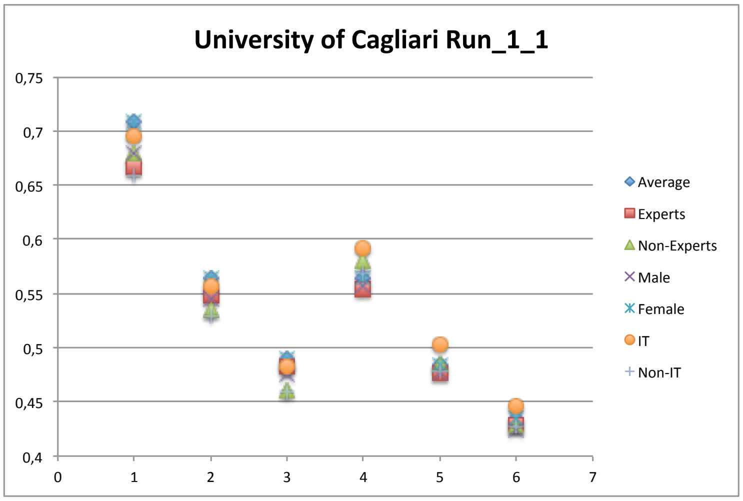


Fig. 5. Retrieval performance for different user groups (University of Cagliari Run1.1)

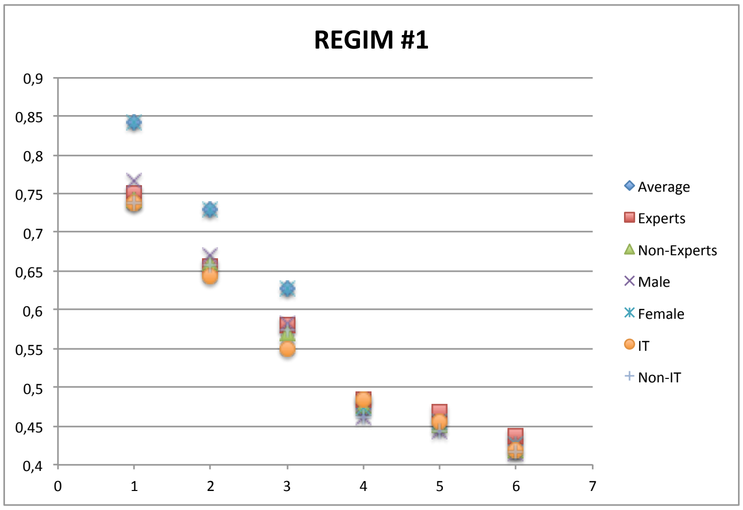


Fig. 6. Retrieval performance for different user groups (REGIM run 1)

References

1. Belkin, N.: Intelligent information retrieval: Whose intelligence? In: ISI '96: Proceedings of the Fifth International Symposium for Information Science. pp. 25–31 (1996)
2. Carterette, B.: System effectiveness, user models, and user utility: a conceptual framework for investigation. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information. pp. 903–912. SIGIR '11, ACM (2011), http://doi.acm.org/10.1145/2009916.2010037
3. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
4. Reiterer, H., Mußler, G., Mann, M.T., Handschuh, S.: INSYDER - an information assistant for business intelligence. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 112–119. SIGIR '00, ACM (2000), http://doi.acm.org/10.1145/345508.345559
5. Voorhees, M.E.: On test collections for adaptive information retrieval. Information Processing & Management 44(6), 1879–1885 (2008), http://www.sciencedirect.com/science/article/pii/S0306457308000253
6. Zellhöfer, D.: An Extensible Personal Photograph Collection for Graded Relevance Assessments and User Simulation. In: Proceedings of the ACM International Conference on Multimedia Retrieval. ICMR '12, ACM (2012)