Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask

David Zellhöfer

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

Abstract. The subtask assesses the retrieval effectiveness in different retrieval usage scenarios in a personal photo collection and with different user groups. That is, the subtask reveals whether a tested algorithm is stable in terms of effectiveness for 7 different user groups. This perspective on retrieval performance evaluation separates the 2013 version of the subtask from its pilot phase although it relies on the same data set. The data set has been sampled from 19 layperson photographers and consists of 5,555 unprocessed digital photographs.

To solve the subtask, the participants are asked to retrieve the 100 best matching documents for 74 sample information needs that consist of visual concepts and events. Each sample information need is modeled by at most one query-by-example document and up to three to browsed documents.

The best performing groups, ISI and DBIS, used visual low-level features and metadata to solve the task. The current best-placed run achieves a nDCG at 20 of 0.7427 for the average user group using relevance feedback and all available modalities, i.e., visual data and metadata such as Exif or GPS information. Regarding the stability, roughly 50% of the submitted runs perform equally well over all user groups.

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

1 Introduction

Following a pilot task phase in 2012, the personal photo retrieval task has become an official subtask of the ImageCLEF photo annotation and retrieval challenge in 2013. The subtask focuses on different retrieval usage scenarios and user groups. That is, the subtask reveals whether the tested algorithms are stable in terms of retrieval quality for different user groups or not. This perspective on retrieval performance evaluation separates the 2013 version of the subtask from its pilot phase [7] although it relies on the same data set. The data set has been sampled from 19 layperson photographers and consists of 5,555 unprocessed digital photographs. A detailed description of the data set is available in [6].

In contrast to system-centric (Cranfield-based) benchmarks, the subtask tries to establish a more user-centered perspective on multimodal information retrieval (MIR) and content-based image retrieval (CBIR). This objective is reflected by three design choices of the subtask.

First, the subtask is not only providing sample information needs (IN) with one or more relevant query documents that have to be used in a query by example (QBE) fashion. In order to simulate the user's interaction with the MIR system, browsed documents are provided in addition to a number of query documents. Unlike the query documents, browsed documents are not necessarily fully relevant for a given topic. Instead, they vary in their level of relevance and can also be totally irrelevant, e.g., to model erroneous user input caused by a click on an image document that has nothing to do with the current IN but that grabbed the user's attention. From a wider perspective, this form of IN specification reflects the transition between different search strategies that have been described, e.g., by [5] or [1].

Second, the subtask respects the gradual relevance of documents with respect to an IN. That is, the subtask's ground truth is based on graded relevance judgments. Consequently, an appropriate metric, nDCG [3], is used for the evaluation of the participants' submission (see Section 4.1).

Third, the subtask acknowledges the subjectivity of relevance assessments. Because user groups that interact with an MIR system and their subjective notion of relevance vary, multiple ground truths are provided for different user groups. Hence, it becomes possible to assess the stability of an examined retrieval algorithm in terms of retrieval effectiveness. This experimental idea was motivated by a preliminary study [7] with the data obtained from the participants of the 2012 pilot task indicating that the algorithms' retrieval performances vary amongst different user groups.

The paper is structured as follows. The next section briefly describes the resources of the subtask, i.e., the data set, the ground truth, and the accompanying baseline system. Section 3 describes the sample information needs (or topics) that are used for the assessment of the retrieval effectiveness of an investigated retrieval algorithm. Section 4 discusses the results of all participants of the subtask, while the last Section 5 concludes the paper.

2 Resources of the Subtask

The subtask relies on the Pythia dataset [6] which equates the data set of the 2012 pilot task on personal photo retrieval that will be described in this section. Hence, the description resembles the publication of 2012 in large parts [7, cf. pp. 1-12]. 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. In addition to the image data, the contributors to the collection completed a survey (see Section 6) 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 subtask. For an overview, see Table 1.

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

To obtain the ground truth, 42 assessors were asked to participate. With the help a web-based evaluation tool (see [7, Fig. 3]), the assessors could judge the relevance of an image with respect to a sample IN (topic) on a graded scale ranging from 0 (irrelevant) to 3 (fully relevant). All assessors had to judge all documents with respect to 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.

Table 2 lists all topics and states whether they belong to an event class or not (see Section 3).

Table 2. Topics of the ImageCLEF 2013 personal photo retrieval subtask

ID	Title	Event	ID	Title	Event	ID	Title
1	Scientific Conference	conference	**26**	Schenna	holiday	**51**	Beach and Seaside
2	Linköping Fire	event	**27**	Umag	holiday	**52**	Street Scene
3	Babelsberg	excursion	**28**	Venice	holiday	**53**	Statue and Figurine
4	Brandenburg	excursion	**29**	Westendorf	holiday	**54**	Asian Temple
5	Eulo	excursion	**30**	Zurich	holiday	**55**	Landscape
6	Sanssouci	excursion	**31**	Die Toten Hosen	rock concert	**58**	Architecture (profane)
7	Telegrafenberg	excursion	**32**	Dream Theater	rock concert	**59**	Animals
8	Flight	flight	**33**	Melt Festival	rock concert	**60**	Asian Temple Interior
9	Altrei	holiday	**34**	Mike Stern	rock concert	**61**	Flower / Botanic Details
10	Bali	holiday	**35**	Toto	rock concert	**63**	Submarine Scene
11	Baltic Sea	holiday	**36**	Transatlantic	rock concert	**64**	Ceremony and Party
12	Cuba	holiday	**37**	U2 1 (Berlin)	rock concert	**65**	Theater / Performing Arts
13	Delft	holiday	**38**	U2 2 (Hannover)	rock concert	**66**	Clouds
14	Dublin	holiday	**39**	Berlin (general)	holiday	**68**	Church (Christian)
15	Edinburgh	holiday	**40**	Cottbus (general)	holiday	**69**	Art Object
16	Grafenau	holiday	**41**	Potsdam (general)	holiday	**70**	Cars
17	Holzleiten	holiday	**42**	Egypt (general)	holiday	**71**	Ship / Maritime Vessel
18	Kleinarl	holiday	**43**	Greece (general)	holiday	**73**	Temple (Ancient)
19	Lenggries	holiday	**44**	Hamburg (general)	holiday	**74**	Squirrels
20	Moscow	holiday	**45**	Mountainside (general)	holiday	**75**	Sign
21	Nassfeld	holiday	**46**	London (general)	holiday	**76**	Mountains
22	New York	holiday	**47**	Party	party	**78**	Birds
23	Padua	holiday	**48**	Rock Concert	rock concert	**79**	Trees
24	Rome	holiday	**49**	Scuba Diving	scuba diving	**81**	City Panorama
25	Scandinavia	holiday	**50**	Soccer			

In order to associate the relevance assessments with different user groups, the assessors had to answer a questionnaire (see Section 6). The questionnaire's outcome is listed in Table 5. The core characteristics of the assessor group 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. 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".

Calculation of the Ground Truths for each Topic Based on the individual assessments, a ground truth for the average user group 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.

In addition to the average user group ground truth, 6 representative user groups could be defined on the basis of the demographics of the assessors that are listed in Table 5. For each of the user groups listed below, a distinct ground truth was derived. In principle, the acquisition of the user group-specific ground truths follows the aforementioned process with the difference that it relies only on relevance assessments that are associated with the specific user group (e.g., expert MIR users). In the event of a missing relevance assessment for the topic-user group combination, the assessment is taken from the average ground truth. The resulting user groups are as follows:

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.
Non-IT The complement of the IT group.

Generation of the 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 0 to 3, i.e., they range from irrelevant to fully relevant for a given topic. In other words, the browsing data consists of interesting images which were not satisfying the information need of the modeled user and motivated him or her to proceed with the search. In contrast to 2012, browsed images could also be irrelevant in order to include erroneous user input.

2.3 Baseline System

In addition to the resources of the 2012 pilot task, the participants were given access to a baseline system that can be used for feature extraction and similarity calculation. The baseline system[1] is available for Linux, Mac OS X, and Windows as C++ source code and is licensed under the Apache License version 2.0. All participants were free to use the system that offers 17 global and local visual features (and some variants).

3 Description of the Sample Information Needs

Unlike in the subtask's pilot phase, the sample information needs (topics) are no longer subdivided into events and visual concepts. An event in the sense of this subtask can be a rock concert or a holiday trip to a region or city. In contrast, a visual concept is a depiction of an object, e.g., a house or a street scene. Table 2 lists all topics including their title and their associated event class (see [6] on the WordNet-based event classes)[2]. The topics 50 to 81 are taken

[1] See http://imageclef.org/2013/photo/retrieval.

[2] Please note that 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].

without modifications from the pilot phase's topic set [7]. The titles were not made available to the participants of the subtask contrasting to 2012 in avoid a manual optimization towards events or visual concepts based on the titles. Additional training data was not released.

For each topic, the sample IN is modeled by at most one fully relevant QBE document and/or a sequence of up to three browsed documents of varying relevance with respect to the IN. 10.81% of the topics (i.e., topics 15, 17, 21, 22, 24, 28, 36, and 42; see Table 2) contain irrelevant browsed documents. The number of topics has been increased to 74 in comparison to 39 during the pilot phase. To summarize, the subtask can be considered more complex in comparison to the pilot, because the IN specification offers less reliable information that can be exploited for query construction.

In consequence, the best matching documents for each topic are expected to be retrieved ad hoc without additional knowledge about the user's context. That is, all participants have to rely on at most one QBE document and/or browsing data and are asked to find the best matching documents illustrating an event or depicting a visual concept. Thus, an additional objective of this task is to find out whether the participating retrieval systems can exploit data from different search strategies, i.e., query-by-example and browsing data, in order to find both visual concepts and photos depicting events. To solve the task, the participants have access to pre-extracted visual low-level features, metadata (e.g. GPS information), but are also free to use their own techniques.

4 Results

In comparison to the pilot phase of the subtask, the participation rate could be increased by ca. 233 %. In 2012, only 3 groups submitted results. This year, 7 groups participated in the subtask, i.e., ca. 38 % of the groups that took part in the ImageCLEF 2013 photo annotation and retrieval challenge. Unfortunately, none of the last year's participants could be motivated to submit runs to the 2013 subtask.

4.1 Evaluation Metrics

As said in the introduction, the relevance of a document with respect to an IN is both highly subjective and relative. That is, a document can be very relevant for an IN while another can be of little value in comparison. To reflect this fact, the presented ground truths are 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, the subtask's retrieval effectiveness evaluation relies on the normalized discounted cumulative gain (nDCG) measurement [3]. As stated in [6] "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. A full discussion of the metric is available by Järvelin and Kekäläinen [3]. For the scope of this task, the DCG implementation of `trec_eval` version 9.0 with standard discount settings is used. For the sake of completeness, MAP at a cut-off level of 100 is also used.

4.2 Results of the Participants

Table 3. Submitted runs and IDs including their type and use of relevance feedback

ID	Run	RF	Type	ID	Run	RF	Type
1	DBIS_run1	None	IMGMETBRO	**14**	ISI_4	None	IMGBRO
2	DBIS_run2	None	IMGMETBRO	**15**	ISI_5	Graded	IMGMETBRO
3	DBIS_run3	Graded	IMGMETBRO	**16**	ThssMpam4_5000_NTI_CR	?	?
4	FINKI_run1	None	IMGBRO	**17**	ThssMpam4_5000_TI_CR	?	?
5	FINKI_run2	None	IMGBRO	**18**	ThssMpam4_5000_TI_NCR	?	?
6	FINKI_run3	None	IMGBRO	**19**	ThssMpam4_5X1000_CR	?	?
7	IPL13_visual_r1	None	IMG	**20**	ThssMpam4_SURFMATCH	?	?
8	IPL13_visual_r2	None	IMG	**21**	VCTLab_1	None	IMGBRO
9	IPL13_visual_r3	None	IMG	**22**	VCTLab_2	None	IMGBRO
10	IPL13_visual_r4	None	IMG	**23**	VCTLab_3	None	IMGBRO
11	ISI_1	Graded	IMGMETBRO	**24**	VCTLab_4	None	IMGBRO
12	ISI_2	Graded	IMGMETBRO	**25**	VCTLab_5	None	IMGBRO
13	ISI_3	Graded	IMGMETBRO	**26**	WideIO	None	IMGBRO

Table 3 lists all participants of the personal photo retrieval subtask including their submitted runs. In total, 7 groups submitted 26 runs. Table 3 shows in addition whether relevance feedback (RF) has been used and which kinds of modalities where exploited during the retrieval. The participants could use the following combinations of the provided document 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)

The best performing groups, ISI and DBIS, used visual low-level features and metadata to solve the task. While ISI used relevance feedback for 4 of their 5 runs, DBIS used this technique only for run #3. Table 4 shows all the results of the different runs ordered by nDCG at cut-off level 20 for the average user group. The complete results for all other user groups are available on the subtask's website[3]. In accordance with the findings of the last years' ImageCLEF tasks, there is evidence that the utilization of multiple modalities increases the retrieval effectiveness.

[3] http://imageclef.org/2013/photo/retrieval#results

The current best-placed run achieves a nDCG at 20 of 0.7427 (average user group) using relevance feedback and all available modalities (IMGMETBRO). In the last year, the best group achieved a nDCG at 20 of 0.5459 for visual concepts (IMGMET, no RF) and a NDCG at 20 of 0.9697 for event retrieval (MET, no RF). Please note that the values values are not meant to be compared directly because of the adjustments made to the subtask in 2013. Unfortunately, none of the former participants could be motivated to submit runs this year. Thus, a statement about an in- or decrease of retrieval effectiveness cannot be made on the basis of the submitted runs. To complicate the matter, only the first two groups have published their algorithms and approaches towards the task. Hence, we cannot provide a complete methodology or retrieval type listing in Table 3. For a description of the methods used by the two first groups, see [4] and [2].

Table 4. Performance of the submitted run for the averaged persona, ordered by nDCG at cut-off level 20

Run	**map_cut_100**	**ndcg_cut_5**	**ndcg_cut_10**	**ndcg_cut_20**	**ndcg_cut_30**	**ndcg_cut_100**
ISI_5	0.5034	0.8104	0.7735	*0.7427*	0.7294	0.6884
ISI_1	0.5028	0.8086	0.7738	*0.7425*	0.7288	0.6878
ISI_2	0.4965	0.8047	0.7633	*0.7379*	0.7271	0.6986
ISI_3	0.4952	0.8057	0.7620	*0.7365*	0.7267	0.6984
DBIS_run3	0.3954	0.7773	0.7197	*0.6798*	0.6546	0.6084
DBIS_run2	0.3767	0.7694	0.7141	*0.6669*	0.6407	0.6082
DBIS_run1	0.3333	0.7516	0.6761	*0.6258*	0.5969	0.5571
ISI_4	0.1855	0.7181	0.6069	*0.5193*	0.4829	0.4236
FINKI_run3	0.1354	0.6878	0.5526	*0.4410*	0.3909	0.3158
FINKI_run2	0.1375	0.6891	0.5510	*0.4398*	0.3881	0.3133
FINKI_run1	0.1360	0.6813	0.5479	*0.4384*	0.3853	0.3109
IPL13_visual_r4	0.1162	0.6627	0.5152	*0.4173*	0.3713	0.3126
IPL13_visual_r1	0.1118	0.6594	0.5152	*0.4125*	0.3725	0.3077
IPL13_visual_r2	0.1082	0.6303	0.4955	*0.3899*	0.3499	0.2910
IPL13_visual_r3	0.0771	0.5769	0.4141	*0.3138*	0.2741	0.2226
ThssMpam4_5000_TI_CR	0.0700	0.5584	0.4005	*0.3051*	0.2676	0.2126
ThssMpam4_5000_TI_NCR	0.0700	0.5572	0.4009	*0.3050*	0.2675	0.2126
VCTLab_2	0.0783	0.4446	0.3574	*0.3047*	0.2754	0.2382
ThssMpam4_5000_NTI_CR	0.0696	0.5606	0.3974	*0.3001*	0.2611	0.2104
ThssMpam4_5X1000_CR	0.0682	0.5547	0.3941	*0.2954*	0.2579	0.2071
VCTLab_1	0.0756	0.4206	0.3420	*0.2950*	0.2731	0.2386
VCTLab_3	0.0751	0.4282	0.3488	*0.2943*	0.2654	0.2336
VCTLab_5	0.0676	0.3816	0.3148	*0.2712*	0.2447	0.2080
VCTLab_4	0.0662	0.3778	0.3128	*0.2616*	0.2412	0.2093
WideIO	0.0584	0.4431	0.3253	*0.2501*	0.2192	0.1845
ThssMpam4_SURFMATCH	0.0529	0.4476	0.3107	*0.2302*	0.1982	0.1494

Retrieval Effectiveness for Different User Groups Figure 1 illustrates the effectiveness variance over the different user groups. The y-axis shows the obtained rank and the x-axis the run ID that is listed in Table 3. Figure 1 shows clearly that roughly 50% of the submitted runs have a low rank variance. That is, they perform equally well for all examined user groups. The other half – predominantly the weak performing runs – is not very stable. Whether this effect correlates with the used features, matching algorithms, or other variables remains an area for further research and cannot be investigated in this paper due to the missing publications.

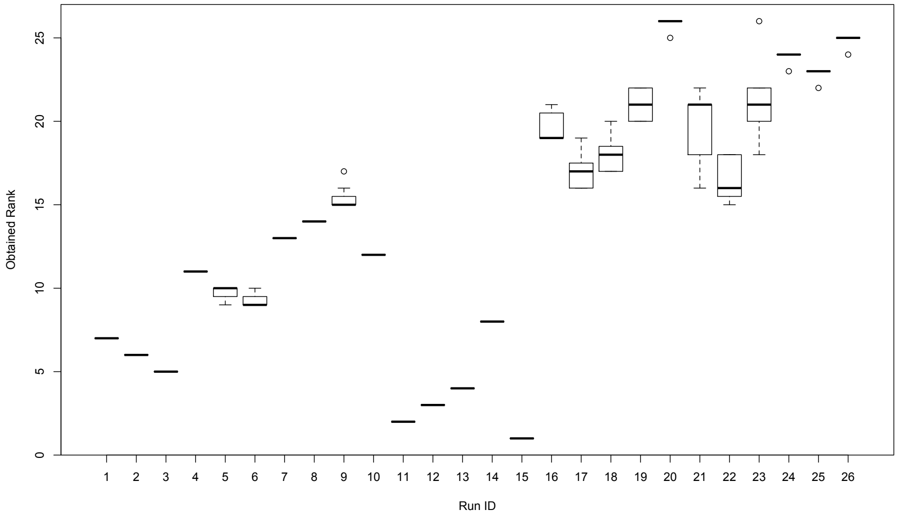

Fig. 1. Obtained ranks over all user groups, see Table 3 for the run IDs

5 Conclusions and Future Work

Although the participation rate in the ImageCLEF 2013 subtask on personal photo retrieval is high, the low publication rate of the participants complicate an interpretation of the results. Anyhow, the results of this subtask strengthen the central finding of the last years of ImageCLEF: the combination of multiple modalities does improve the retrieval effectiveness.

The interpretation of the stability of the submitted runs indicates that there might be a correlation between the effectivity and stability of an algorithm. In other words, the better one's algorithm performs the more likely it is that it will do so for different user groups. Whether this effect is due to other (hidden) variables remains an open question. Maybe this question motivates the missing participants to publish their algorithms and approaches towards the solution of the subtask. Because of the low publication rate, a general interpretation of the results is hardly possible.

Another interesting result of the conducted experiment is that both leading groups – ISI and DBIS – perform almost equally well although ISI is relying on sophisticated techniques such as Fisher vectors and local features while DBIS uses global low-end features embedded in a logical query language. Given the fact, that local features are computationally more intensive than global features, one might further investigate the logical combination of global features in order to achieve comparable results at less computational costs.

6 Content of the Usage Questionnaire

- **Year of Birth**
- **Gender**
- **Job Type** 1) Pupil, 2) In job training, 3) Student, 4) Fully employed, 5) Part-time employed, 6) Not employed, 7) Retired, 8) Other
- **Field of Study / Job Training**
- **Course Level**

Q0: Have you visited one or more oft he following lectures?
IR) Information Retrieval, MR) Multimedia Retrieval

Q1: Are you familiar with the principles of content-based information retrieval?
0) No, 1) A little, 2) I am an informed outsider, 3) Very much, 4) I am an expert

Q2: Are you colorblind? 0) I don't know, 1) No, 2) Yes

Q3: How many minutes do you use the internet per day?
0) Not at all, 2) 1 - 30 minutes, 3) 31 - 60 minutes. 4) 61 - 90 minutes, 5) 91 - 120 minutes, 6) More than 120 minutes, 7) More than 240 minutes

Q4: Do you know Web 2.0 services such as Flickr or Fotocommunity.de for sharing holiday, family or other photographs with friends?
0) Never heard of it, 2) Know it by name, 3) I have visited such websites, 4) I do have an account

Q5: How often do you use such Web 2.0 services to share photographs with friends? 0) Never, 1) Less than once a month, 2) More than once a month, 3) Weekly, 4) Daily

Q6: Which of the following services do you use to upload and administrate holiday, family or other photographs? (Choose one or more.)
None, Facebook, Flickr, Fotocommunity.de, Picasa, Other

Q7: How often do you take photographs?
0) Seldom, 1) Only at special events, 2) Often, 3) Virtually always

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. Böttcher, T., Zellhöfer, D., Schmitt, I.: BTU DBIS' Personal Photo Retrieval Runs at ImageCLEF 2013. In: CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain (2013)
3. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
4. Mizuochi, M., Higuchi, T., Kamada, C., Harada, T.: MIL at ImageCLEF 2013: Personal Photo Retrieval. In: CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain (2013)

Table 5. Demographics of the assessors, for the answer codes see Section 6

AssessorID	Year of Birth	Gender	Job Type	Field of Study / Job Training	Course Level	MR	IR	Q1	Q2	Q3	Q4	Q5	Q7	Q6	Q6	Q6
assessor11	1988	m	3	Business Administration	M.Sc.	0	0	0	0	3	3	1	2	facebook	flickr	
assessor18	1985	m	3	Business Administration	B.Sc.	0	0	0	1	6	1	1	2	facebook		
assessor24	1990	f	3	Business Administration	B.Sc.	0	0	0	1	5	3	2	1	facebook	picasa	
assessor27	1987	f	3	Business Administration	B.Sc.	0	0	0	1	6	1	1	3	facebook		
assessor48	1983	f	3	Business Administration	PhD	0	0	0	0	6	2	0	1			
assessor13	1988	m	3	Business Administration	M.Sc.	0	0	1	1	5	2	0	2			
assessor28	1988	f	3	Business Administration	M.Sc.	0	0	1	1	3	0	1	2	facebook	picasa	
assessor26	1987	f	3	Business Administration	M.Sc.	0	0	2	1	3	0	0	1	picasa		
assessor25	1984	f	3	Business Administration	B.Sc.	0	0	3	1	3	0	0	2			
assessor10	1988	m	3	Business Administration & Engineering	M.Sc.	0	0	0	0	4	3	1	2	fotocommunity		
assessor12	1988	m	3	Business Administration & Engineering	B.Sc.	0	0	0	0	6	0	1	2	facebook		
assessor16	1986	m	3	Business Administration & Engineering	M.Sc.	0	0	0	1	5	1	0	2	facebook		
assessor17	1991	m	3	Business Administration & Engineering	B.Sc.	0	0	0	1	5	2	0	1			
assessor30	1988	m	3	Business Administration & Engineering	M.Sc.	0	0	0	1	5	3	0	3			
assessor21	1986	m	3	Business Administration & Engineering	M.Sc.	0	1	1	1	3	2	2	1			
assessor31	1989	m	3	Business Administration & Engineering	M.Sc.	0	0	1	1	5	3	1	2	facebook	flickr	other
assessor33	1986	m	3	Business Administration & Engineering	M.Sc.	0	1	1	1	6	0	0	0			
assessor22	1988	f	3	Business Administration & Engineering	M.Sc.	0	0	1	1	5	1	1	1	facebook		
assessor23	1987	f	3	Business Administration & Engineering	M.Sc.	0	0	2	1	2	1	0	1			
assessor20	1987	m	3	Computer Science	M.Sc.	0	0	0	1	6	2	0	0			
assessor51	1990	m	3	Computer Science	B.Sc.	0	0	0	1	5	2	0	0			
assessor37	1987	m	3	Computer Science	M.Sc.	1	1	2	1	6	0	0	2			
assessor36	1986	m	3	Computer Science	M.Sc.	1	1	3	1	3	2	0	1			
assessor41	1988	f	3	Computer Science	M.Sc.	1	1	3	1	6	1	0	1			
assessor42	1985	f	4	Computer Science		1	0	3	1	5	1	0	1			
assessor2	1979	m	4	Computer Science	PhD	1	1	4	1	6	3	1				
assessor44	1981	m	4	Computer Science		0	1	4	0	6	2	0	1			
assessor50	1985	m	4	eBusiness		0	0	0	1	6	2	1	1			
assessor32	1987	m	3	eBusiness	M.Sc.	1	1	1	1	4	2	0	1			
assessor39	1981	m	3	eBusiness	M.Sc.	0	0	1	0	6	1	0	2			
assessor29	1988	f	3	eBusiness	M.Sc.	0	0	1	1	5	1	0	0			
assessor38	1991	f	3	eBusiness	B.Sc.	0	0	1	1	3	1	0	1			
assessor46	1982	f	4	eBusiness	PhD	0	0	2	1	6	2	1	1	other		
assessor53	1988	m	3	eBusiness	M.Sc.	1	1	3	0	6	2	2	1	facebook		
assessor15	1989	m	3	Information & Media Technology	M.Sc.	0	0	1	1	6	3	1	0	facebook	flickr	
assessor19	1985	m	3	Information & Media Technology	M.Sc.	0	0	1	1	4	2	1	0	flickr		
assessor35	1986	m	3	Information & Media Technology	M.Sc.	0	1	2	1	5	1	0	1			
assessor45	1982	m	4	Information & Media Technology		1	1	3	1	5	1	0	1			
assessor43	1984	m	4	Information & Media Technology		1	0	4	1	3	3	1	1	picasa		
assessor14	1987	m	3	Urban & Regional Planning	M.Sc.	0	0	1	1	5	3	1	1			
assessor49	1985	m	4			0	0	3	1	6	3	1	1	facebook		
assessor47	1984	f	4			0	0	3	1	6	2	1	1	other		
Min	1979	**Min**	3				**Min**	0	0	2	0	0	0			
Max	1991	**Max**	4				**Max**	4	1	6	3	2	3			
Median	1987	**Median**	3				**Median**	1	1	5	2	0	1			
Mean	1986.29	**Mean**	3.21				**Mean**	1.40	0.83	4.88	1.67	0.55	1.22			
	Male	28			**Non-visited class**	33	31									
	Female	14			**Visited class**	9	11									

5. 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
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)
7. Zellhöfer, D.: Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop (2012)