Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009

Monica Lestari Paramita, Mark Sanderson and Paul Clough
{m.paramita, m.sanderson, p.d.clough}@sheffield.ac.uk
University of Sheffield, United Kingdom

Abstract

The ImageCLEF Photo Retrieval Task 2009 focused on image retrieval and diversity. A new collection was utilised in this task consisting of approximately half a million images with English annotations. Queries were based on analysing search query logs and two different types were released: one containing information about image clusters; the other without. A total of 19 participants submitted 84 runs. Evaluation, based on Precision at rank 10 and Cluster Recall at rank 10, showed that participants were able to generate runs of high diversity and relevance. Findings show that submissions based on using mixed modalities performed best compared to those using only concept-based or content-based retrieval methods. The selection of query fields was also shown to affect retrieval performance. Submissions not using the cluster information performed worse with respect to diversity than those using this information. This paper summarises the ImageCLEFPhoto task for 2009.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages-*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Performance Evaluation, Image Retrieval, Diversity, Clustering

1 Introduction

The ImageCLEFPhoto task is part of the CLEF evaluation campaign, the focus for the past two years being promoting diversity within image retrieval. The task originally began in 2003 and has since attracted participants from many institutions worldwide. For the past three years, ImageCLEFPhoto has used a dataset of 20,000 general photos called the IAPR TC-12 Benchmark. In 2008, we adapted this collection to enable the evaluation of diversity in image retrieval results. We recognised that this setup had limitations and therefore moved to using a larger and more realistic collection of photos (and associated search query logs) from Belga[1], a Belgian press agency. Even though photos in this collection have English-only annotations and hence provide little challenge to cross-language information retrieval systems, there are other characteristics of the dataset which provide new challenges to participating groups (explained in Section 1.1). The resources created for the 2009 task have given us the opportunity to study diversity for image retrieval in more depth.

1.1 Evaluation Scenario

Given a set of information needs (topics), participants were tasked with finding not only relevant images, but also generating ranked lists that promote diversity. To make the task harder, we released two types of queries: the first type of query included written information about the specific requirement for diversity (represented as *clusters*); queries of the second type contained a more conventional title and example relevant images. In the former type of query participants were required to retrieve diverse results with some indication of what types of clusters were being sought; in the latter type of query little evidence was given for what kind of diversity was required. Evaluation gave more credence to runs that presented diverse results without sacrificing precision than those exhibiting less diversity.

[1] Belga Press Agency: http://www.belga.be

1.2 Evaluation Objectives for 2009

The Photo Retrieval task in 2009 was focused at studying diversity further. Using resources from Belga, we provided a much larger collection, containing just under half a million images, compared to 20,000 images provided in 2008. We also obtained statistics on popular queries submitted to the Belga website in 2008 [1], which we exploited to create representative queries for this diversity task. We experimented with different ways of specifying the need for diversity which was given to participants, and this year decided to release half of the queries without any indication of diversity required or expected. We were interested in addressing the following research questions:

- Can results be diverse without sacrificing relevance?
- How much will knowing about query clusters a priori help increase diversity in image search results?
- Which approaches should be used to maximize diversity and relevance for image search results?

These research questions will be discussed further in section 4.

2 Evaluation Framework

One of the major challenges for participants of the 2009 ImageCLEFPhoto task was a new collection which was 25 times larger than that used for 2008. Query creation was based completely on query log data, which helped to make the retrieval scenario as realistic as possible [2]. We believe this new collection will provide a framework in which to conduct a more thorough analysis of diversity in image retrieval.

2.1 Document Collection

The collection consists of 498,920 images with English-only annotations (i.e. captions) describing the content of the image. However, different to the structured annotations of 2008, the annotations in this collection are presented in an unstructured way (Table 1). This increases the challenge for participants as they must automatically extract information about the location, date, photographic source, etc of the image as a part of the indexing and retrieval process. The photos cover a wide-ranging time period, and there are many cases where pictures have not been orientated correctly, thereby increasing the challenge for content-based retrieval methods.

Table 1. Example image and caption

	Annotation: 20090126 - DENDERMONDE, BELGIUM: Lots of people pictured during a commemoration for the victims of the knife attack in Sint-Gilles, Dendermonde, Belgium, on Monday 26 January 2009. Last friday 20-Year old Kim De Gelder killed three people, one adult and two childs, in a knife attack at the children's day care center "Fabeltjesland" in Dendermonde. BELGA PHOTO BENOIT DOPPAGNE

2.2 Query Topics

Based on search query logs from Belga, 50 example topics were generated and released as two query types (as mentioned previously). From this set, we randomly chose 25 queries to be released with information including the title, cluster title, cluster description and image (example) as shown in Table 2. We refer to these queries as *Query Part 1*. In this example, participants can notice that this result about 'Clinton' requires 3 different clusters, which are 'Hillary Clinton', 'Obama Clinton' and 'Bill Clinton'. Results covering other aspects of "Clinton", such as Chelsea Clinton or Clinton Cards, will not be counted towards the final diversity score. More information about these clusters and the method used to produce them can be found in [2].

Given that one might argue that the diversity result in *Query Part 1* could be relatively easy to produce as detailed information about the different sub-topics is provided as part of the query topic and there are often in practice instances when little or no query log information is available to indicate possible clusters, we released 25 queries containing no information about the kind of diversity expected (referred to as *Query Part 2*). An

example of this query type is given in Table 3. It should be noted that information about the cluster titles and description were also based on Belga's query logs. However, we did not release any of this information to the participants.

Table 2. Example of *Query Part 1*

```
<top>
<num> 12 </num>
<title> clinton </title>
<clusterTitle> hillary clinton </clusterTitle>
<clusterDesc> Relevant images show photographs of Hillary Clinton. Images
of Hillary with other people are relevant if she is shown in the
foreground. Images of her in the background are irrelevant. </clusterDesc>
<image> belga26/05859430.jpg </image>
<clusterTitle> obama clinton </clusterTitle>
<clusterDesc> Relevant images show photographs of Obama and Clinton. Images
of those two with other people are relevant if they are shown in the
foreground. Images of them in the background are irrelevant. </clusterDesc>
<image> belga28/06019914.jpg </image>
<clusterTitle> bill clinton </clusterTitle>
<clusterDesc> Relevant images show photographs of Bill Clinton. Images of
Bill with other people are relevant if he is shown in the foreground.
Images of him in the background are irrelevant. </clusterDesc>
<image> belga44/00085275.jpg </image>
</top>
```

Table 3. Example of *Query Part 2*

```
<top>
<num> 26 </num>
<title> obama </title>
<image> belga30/06098170.jpg </image>
<image> belga28/06019914.jpg </image>
<image> belga30/06107499.jpg </image>
</top>
```

The list of 50 topics used in this collection is given in Table 4. Since Belga is a press agency based in Belgium, there are a large number of queries which contain the names of Belgian politicians, Belgian football clubs and members of the Belgian royal family. Other queries, however, are more general such as Beckham, Obama, etc. There are some queries which are very broad and under-specified (e.g. Belgium); others are highly ambiguous (e.g. Prince and Euro).

Table 4. Overall list of topics used in the 2009 task

Query Part 1				Query Part 2			
1	leterme	**14**	**princess****	26	obama*	**39**	**beckham***
2	fortis	15	monaco**	**27**	**anderlecht**	**40**	**prince****
3	**brussels****	**16**	**queen****	**28**	**mathilde**	41	princess mathilde
4	**belgium****	17	tom boonen	29	boonen	42	mika*
5	**charleroi**	**18**	**bulgaria****	**30**	**china****	43	ellen degeneres
6	vandeurzen	**19**	**kim clijsters**	31	hellebaut	**44**	**henin**
7	gevaert	**20**	**standard**	32	nadal	**45**	**arsenal**
8	koekelberg	21	princess maxima	**33**	**snow****	**46**	**tennis****
9	daerden	22	club brugge	**34**	**spain****	47	ronaldo*
10	borlee*	**23**	**royals****	35	strike**	48	king**
11	**olympic****	**24**	**paola***	**36**	**euro***	49	madonna
12	**clinton***	25	mary*	**37**	**paris****	**50**	**chelsea**
13	martens*			38	rochus		

* = ambiguous, ** = under-specified queries, **bold queries**: queries with more than 677 (median) relevant documents

2.3 Relevance Assessments

Relevance assessments were performed using the DIRECT (Distributed Information Retrieval Evaluation Campaign Tool)[2], a system which enables assessors to work in a collaborative environment. We hired 25 assessors to be involved in this process and assessments were divided into 2 phases: in the first phase, assessors were asked to identify images relevant to a given query. Information about all relevant clusters to the topic was given to assessors to ensure they were aware of the scope of relevant images for a query. The number of relevant images for each query resulting from this stage is shown in Figure 1.

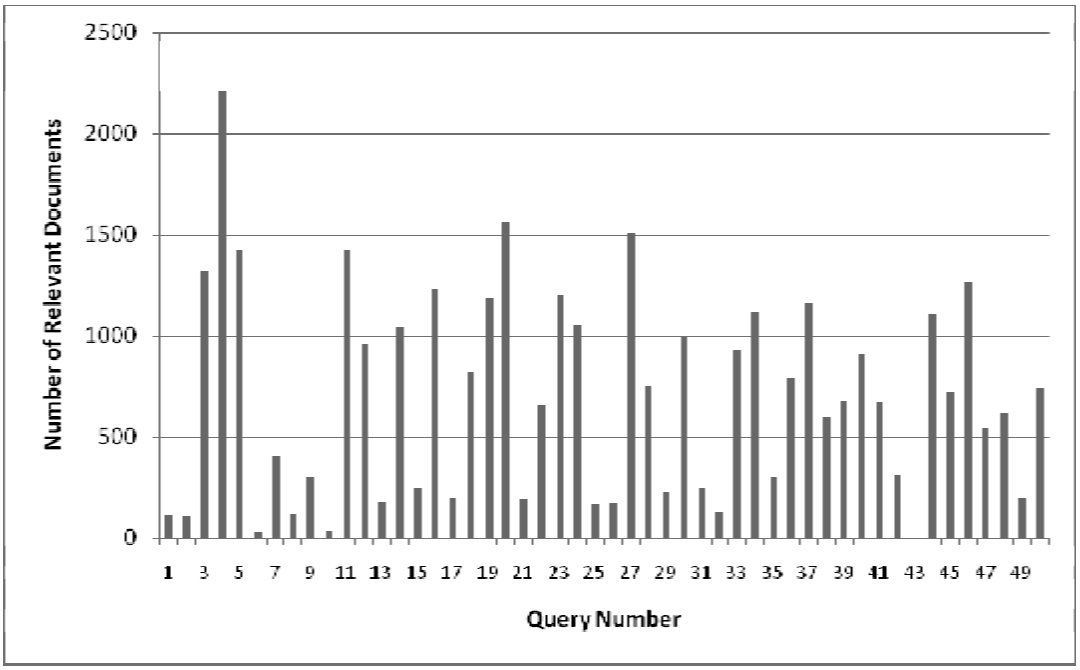

Figure 1. Number of relevant documents per query

Having queries from different types shown in Table 4, we then analysed the number of relevant documents in each type. This data, shown in Table 5, illustrates that under specified queries have the highest average number of relevant documents.

Table 5. Number of relevant documents in each type

	All Queries	Ambiguous Queries	Under Specified Queries	Other Queries
Number of Queries	50	10	16	24
Average Doc	697.74	490	*1050.19*	549.33
Min	2	35	246	*2*
Max	2210	1052	*2210*	1563
Standard Dev	512.16	366.28	459.29	490.5

After a set of relevant images were found, for the second stage different assessors were asked to find images relevant to each cluster (some images could belong to multiple clusters). Since topics varied widely in content and diversity, the number of relevant images varied from 1 to 1,266 for each cluster. Initially, there were 206 clusters created for the 50 queries, but this number dropped to 198 as there were 8 clusters with no relevant images which had to be deleted. There are an average number of 208.49 relevant documents for each cluster, with a standard deviation of 280.59. The distribution of clusters is shown in Figure 2.

[2] http://direct.dei.unipd.it

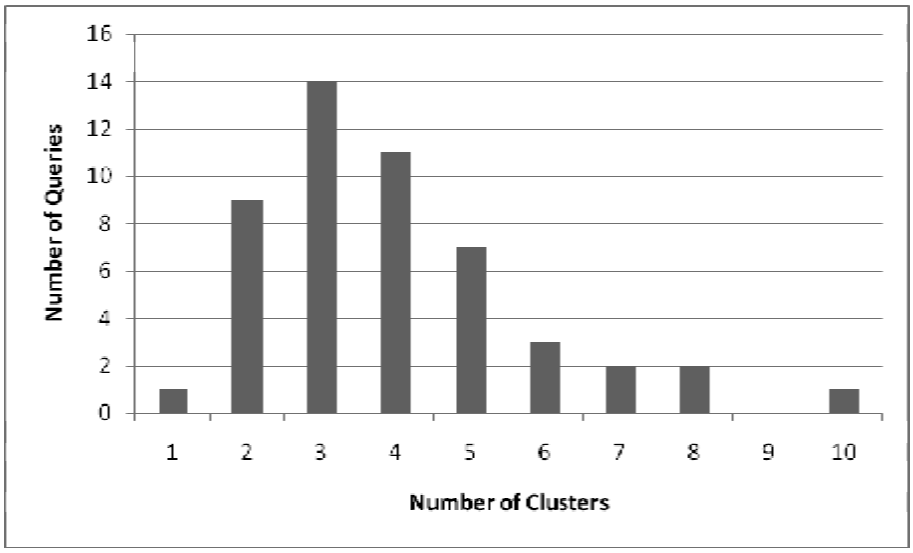

Figure 2. Distribution of clusters in the queries

2.4 Generating the Results

The method for generating results from participant's submissions was similar to that used in 2008 [3]. The precision of each run (P@10) was evaluated using *trec_eval* and cluster recall (CR@10) was used to measure diversity. Since the maximum number of clusters was set to 10 [2], we focussed evaluation on P@10 and CR@10. The F_1 score calculates the harmonic mean of these two measures.

3 Overview of Participation and Submissions

A total of 44 different institutions registered for the ImageCLEFPhoto task (the highest number of applications ever received for this task). From this number, 19 institutions from 10 different countries finally submitted runs to the evaluation. Due to the large number of runs received last year, we limited the number of submitted runs to 5 per participant. A total of 84 runs were submitted and evaluated (some groups submitted less than 5 runs).

3.1 Overview of Submissions

The participating groups for 2009 are listed in Table 8. From the 24 groups participating in the 2008 task, 15 groups returned and were involved this year (Returning). We also received four new participants who joined this task for the first time (New).

Participants were asked to specify the query fields used in their search and the modality of the runs. Query fields were described as T (Title), CT (Cluster Title), CD (Cluster Description) and I (Image). The modality was described as TXT (text-based search only), IMG (content-based image search only) or TXT-IMG (both text and content-based image search). The range of approaches is shown in Tables 6 and 7 and summarised in Figure 3.

Table 6. Choice of query fields

Query Fields	Number of Runs
T	17
T-CT-CD-I	15
T-CT	15
T-CT-I	9
T-CT-CD	9
I	8
T-I	7
CT-I	2
CT	2

Table 7. Modality of the runs

Modality	TXT-IMG	TXT	IMG
Number of Runs	36	41	7

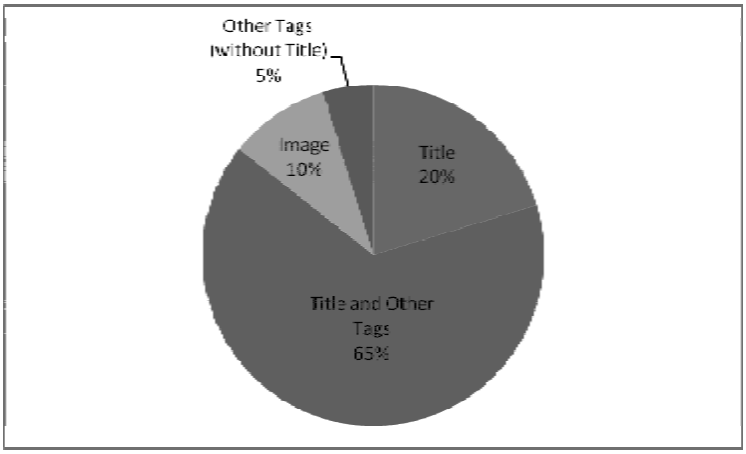

Figure 3. Summary of query fields used in submitted runs

Table 8. Participating groups

No	Group ID	Institution	Country	Runs	Status
1	Alicante	University of Alicante	Spain	5	Returning
2	Budapest-ACAD	Hungarian Academy of Science, Budapest	Hungary	5	Returning
3	Chemnitz	Computer Science, Trinity College, Dublin	Ireland	4	Returning
4	CLAC-Lab	Computational Linguistics at Concordia (CLAC) Lab, Concordia University, Montreal	Canada	4	Returning
5	CWI	Interactive Information Access	Netherlands	5	New
6	Daedalus	Computer Science Faculty, Daedalus, Madrid	Spain	5	Returning
7	Glasgow	Multimedia IR, University of Glasgow	UK	5	Returning
8	Grenoble	Lab. Informatique Grenoble	France	4	Returning
9	INAOE	Language Tech	Mexico	5	Returning
10	InfoComm	Institution for InfoComm Research	Singapore	5	Returning
11	INRIA	LEAR Team	France	5	New
12	Jaen	Intelligent Systems, University of Jaen	Spain	4	Returning
13	Miracle-GSI	Intelligent System Group, Daedalus, Madrid	Spain	3	Returning
14	Ottawa	NLP, AI.I.Cuza U. of IASI	Canada	5	Returning
15	Southampton	Electronics and Computer Science, University of Southampton	UK	4	New
16	UPMC-LIP6	Department of Computer Science, Laboratoire d'Informatique de Paris 6	France	5	Returning
17	USTV-LSIS	System and Information Sciences Lab, France	France	2	Returning
18	Wroclaw	Wroclaw University of Technology	Poland	5	New
19	XEROX-SAS	XEROX Research	France	4	Returning

4 Results

This section provides an overview of the results based on the type of queries and modalities used to generate the runs. As mentioned in the previous section, we used P@10 to calculate the fraction of relevant documents in the top 10 and CR@10 to evaluate diversity, which calculates the proportion of subtopics retrieved in the top 10 documents as shown below:

$$Cluster-recall\ at\ K \equiv \frac{\left|\bigcup_{i=1}^{K} subtopics(d_i)\right|}{n_A}$$

The F_1 score was used to calculate the harmonic mean of P@10 and CR@10, to enable the results to be sorted by one single measure:

$$F_1 = \frac{|\ 2\ x\ (P10\ x\ CR10)\ |}{(P10 + CR10)}$$

4.1 Results across all Queries

The top 10 runs computed across all 50 queries (ranked in descending order of F_1 score) are shown in Table 9.

Table 9. Systems with highest F_1 score for all queries

No	Group	Run Name	Query	Modality	P@10	CR@10	F_1
1	XEROX-SAS	XRCEXKNND	T-CT-I	TXT-IMG	0.794	0.8239	**0.8087**
2	XEROX-SAS	XRCECLUST	T-CT-I	TXT-IMG	0.772	0.8177	**0.7942**
3	XEROX-SAS	KNND	T-CT-I	TXT-IMG	0.8	0.7273	**0.7619**
4	INRIA	LEAR5_TI_TXTIMG	T-I	TXT-IMG	0.798	0.7289	**0.7619**
5	INRIA	LEAR1_TI_TXTIMG	T-I	TXT-IMG	0.776	0.7409	**0.7580**
6	InfoComm	LRI2R_TI_TXT	T-I	TXT	0.848	0.6710	**0.7492**
7	XEROX-SAS	XRCE1	T-CT-I	TXT-IMG	0.78	0.7110	**0.7439**
8	INRIA	LEAR2_TI_TXTIMG	T-I	TXT-IMG	0.772	0.7055	**0.7373**
9	Southampton	SOTON2_T_CT_TXT	T-CT	TXT	0.824	0.6544	**0.7294**
10	Southampton	SOTON2_T_CT_TXT_IMG	T-CT	TXT-IMG	0.746	0.7095	**0.7273**

Looking at the top 10 runs, we observe that highest effectiveness is reached using mixed modality (text and image) and using information from the query title, cluster title and the image content itself. The scores for P@10, CR@10 and F_1 in this year's task are notably higher than the evaluation last year. Moreover, the number of relevant images in this year's task was higher. Having two different types of queries, we analysed how participants dealt with the different queries. Tables 10 and 11 summarise the top 10 runs in each of query types.

Table 10. Systems with highest F_1 score for *Queries Part 1*

No	Group	Run Name	Query	Modality	P@10	CR@10	F_1
1	Southampton	SOTON2_T_CT_TXT	T-CT	TXT	0.868	0.7730	**0.8178**
2	Southampton	SOTON2_T_CT_TXT_IMG	T-CT	TXT-IMG	0.804	0.8063	**0.8052**
3	XEROX-SAS	KNND	T-CT-I	TXT-IMG	0.768	0.8289	**0.7973**
4	XEROX-SAS	XRCE1	T-CT-I	TXT-IMG	0.768	0.8289	**0.7973**
5	XEROX-SAS	XRCECLUST	T-CT-I	TXT-IMG	0.768	0.8289	**0.7973**
6	XEROX-SAS	XRCEXKNND	T-CT-I	TXT-IMG	0.768	0.8289	**0.7973**
7	Southampton	SOTON1_T_CT_TXT	T-CT	TXT	0.824	0.7470	**0.7836**
8	InfoComm	LRI2R_TCT_TXT	T-CT	TXT	0.828	0.7329	**0.7776**
9	Southampton	SOTON1_T_CT_TXT_IMG	T-CT	TXT-IMG	0.76	0.7933	**0.7763**
10	INRIA	LEAR1_TI_TXTIMG	T-I	TXT-IMG	0.772	0.7779	**0.7749**

Different compared to results presented previously, it is interesting to see that the top run in *Queries Part 1* used only text retrieval approaches. Even though the CR@10 score was lower than most of the runs, it obtained the highest F_1 score due to a high P@10 score. The uses of tags vary within results, but the top 9 runs consistently

use both title and cluster title. We therefore conclude that the use of title and cluster title do help the participants to achieve a good score in both precision and cluster recall.

In the queries part two, participants did not have access to cluster information. We specifically intended this to see how well the system finds diverse results without any hints. The results of the top runs in queries part 2 is shown in Table 11.

Table 11. Systems with highest F_1 score for *Queries Part 2*

No	Group	Run Name	Query	Modality	P@10	CR@10	F_1
1	XEROX-SAS	XRCEXKNND	T-I	TXT-IMG	0.82	0.8189	**0.8194**
2	XEROX-SAS	XRCECLUST	T-I	TXT-IMG	0.776	0.8066	**0.7910**
3	InfoComm	LRI2R_TI_TXT	T-I	TXT	0.828	0.6901	**0.7528**
4	INRIA	LEAR5_TI_TXTIMG	T-I	TXT-IMG	0.756	0.7399	**0.7479**
5	INRIA	LEAR1_TI_TXTIMG	T-I	TXT-IMG	0.78	0.7039	**0.7400**
6	GRENOBLE	LIG3_TI_TXTIMG*	T-I	TXT-IMG	0.7708	0.6711	**0.7175**
7	XEROX-SAS	KNND	T-I	TXT-IMG	0.832	0.6257	**0.7143**
8	INRIA	LEAR2_TI_TXTIMG	T-I	TXT-IMG	0.728	0.6849	**0.7058**
9	GRENOBLE	LIG4_TCTITXTIMG	T-I	TXT-IMG	0.792	0.6268	**0.6998**
10	GLASGOW	GLASGOW4	T	TXT	0.76	0.6401	**0.6949**

** submitted results for 24 out of 25 queries. Score shown is the average of the submitted queries only.*

It is shown in the table that the top 9 runs use information from example images, which shows that example images and their annotations might have given useful hints to detect diversity. To analyse this further, we divided the runs which used the Image field and those which did not, and found that the average CR@10 scores were 0.5571 and 0.5270 respectively. We conclude that having example images helps to identify diversity and present a more diverse set of results.

Comparing the CR@10 scores in the top 10 runs of *Queries Part 1* and *Queries Part 2*, the scores in the latter group were lower, which implied that systems did not find as many diverse results when cluster information was not available. The F_1 scores from these top 10 were also lower, but they only differed slightly compared to the *Queries Part 1*. We also calculated the magnitude of difference between results for different query types (shown in Table 12). This indicates that on average runs do perform lower in *Query Part 2*, however the difference is small and not sufficient to conclude that runs will be less diverse if cluster titles are not available (p=0.146).

Table 12. Cluster Recall score difference between *Queries Part 1* and *Queries Part 2*

Mean	StDev	Max	Min
-0.0234	0.1454	0.2893	-0.6459

It is important to understand that not all the runs in *Query Part 1* use the cluster title. To analyse how useful the "Cluster Title" (CT) information is, we divided the runs of *Query Part 1* based on the use of CT field. The mean and standard deviation of P@10, CR@10 and the F_1 scores is shown in Table 13 (the highest score shown in italics).

Table 13. Comparison of CR@10 scores

Queries	Number of Runs	P@10		CR@10		F_1	
		Mean	SD	Mean	SD	Mean	SD
Query part 1 with CT	52	*0.6845*	0.2	*0.5939*	0.1592	*0.6249*	0.1701
Query part 1 without CT	32	0.6641	0.2539	0.5006	0.1574	0.5581	0.1962
Query part 2	84	0.6315	0.2185	0.5415	0.1334	0.5693	0.1729

Table 13 provides more evidence that the Cluster Title field has an important role in identifying diversity. When Cluster Title is not being used, the F_1 scores of both *Query Part 1* and *Query Part 2* do not differ significantly. Figure 3 shows a scatter plot of F_1 scores for each query type. Using a two-tailed paired t-test, the scores between *Queries Part 1* and *Queries Part 2* were found to be significantly different (p=0.02). There is also a significant correlation between the scores: the Pearson correlation coefficient equals 0.691.

We evaluated the same test on the runs using Cluster Title only to the runs in *Query Part 2*, and found that they are also significantly different (p=0.003), the Pearson correlation coefficient equals 0.745. However, when the same evaluation was being performed on runs not using Cluster Title, the difference in scores was not significant (p=0.053), although obtaining a Pearson correlation coefficient of 0.963.

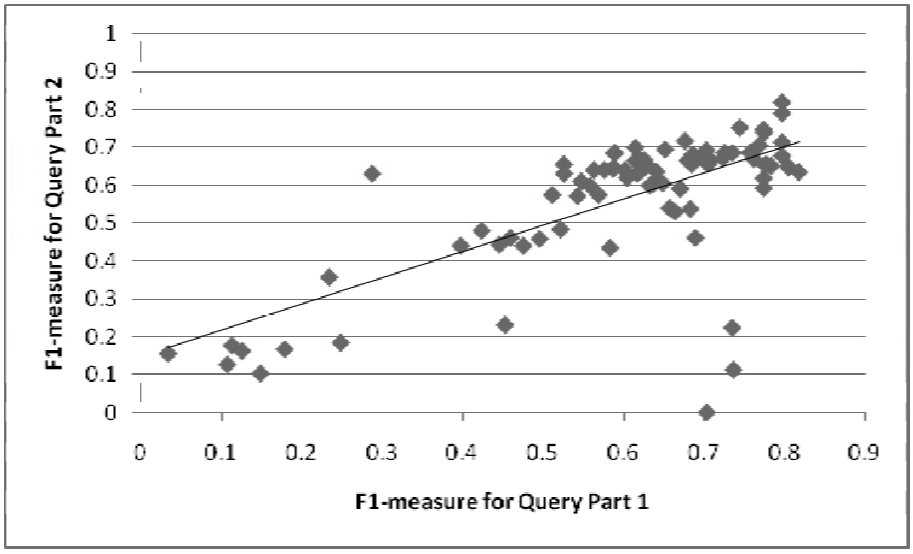

Figure 4. Scatter plot for F_1 scores of each run between query types

Table 14 summarises the results across all queries (mean scores). According to these results, highest scores from the three conditions are obtained when the query has full information about potential diversity.

Table 14. Summary of results across all queries

Queries	P@10		CR@10		F_1	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
All Queries	0.655	0.2088	0.5467	0.1368	0.5848	0.1659
Query Part 1	*0.6768*	0.2208	*0.5583*	0.1641	*0.5995*	0.1823
Query Part 2	0.6315	0.2185	0.5415	0.1334	0.5693	0.1729

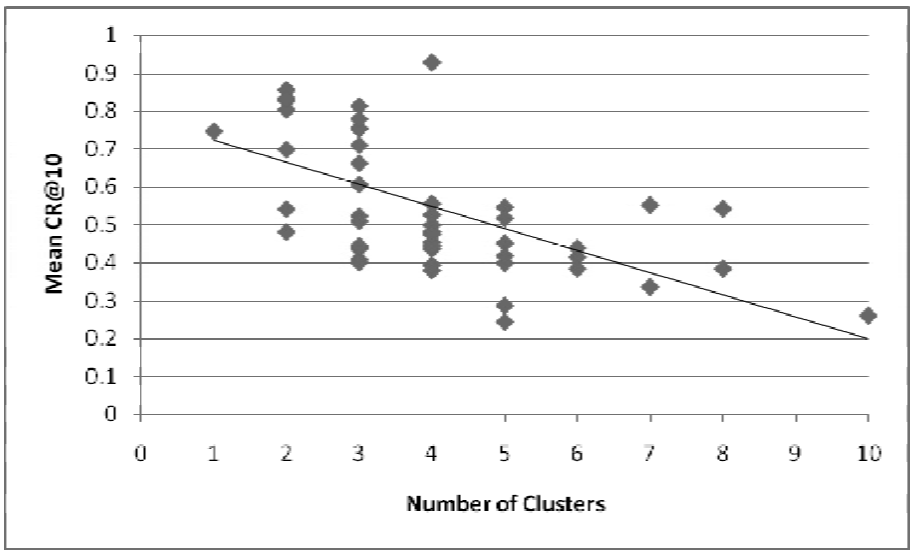

Figure 5. Scatter plot for mean CR@10 scores for each query

We also analysed whether the number of clusters have any effect on the diversity score. To measure this factor, we calculated the mean CR@10 for all of the runs. These scores are then plotted based on the number of clusters contained in each specified query. This scatter plot, shown in Figure 5, has a Pearson correlation coefficient of -0.600, confirming that the more clusters a query contains, the lower the CR@10 score is.

4.2 Results by Retrieval Modality

In this section, we will present an overview result of runs using different modalities.

Table 15. Results by retrieval modality

Modality	Number of Runs	P@10		CR@10		F_1	
		Mean	SD	Mean	SD	Mean	SD
TXT-IMG	36	*0.713*	0.1161	*0.6122*	0.1071	*0.6556*	0.1024
TXT	41	0.698	0.142	0.5393	0.0942	0.5976	0.0964
IMG	7	0.103	0.027	0.2535	0.0794	0.1456	0.0401

According to Table 15, both the precision and cluster recall scores are highest if systems use both low-level features based on the content of an image and its associated text. The mean of the runs using image content only (IMG) is drastically lower based on the P@10 score; however the gap decreases when considering only the CR@10 score. Further research should be carried out to improve runs using content-based approaches only, as the best run using this approach had the lowest F_1 score (0.218) compared to TXT (0.351) and TXT-IMG (0.297).

4.3 Approaches Used by Participants

Having known that the mixed modality performs best, we were also interested to see the best combination of query fields to maximize the F_1 score of the runs. We therefore calculated the mean of each combination and modality and the result is shown in Table 16 with the highest score for each modality shown in italic.

Table 16. Choice of query tags with mean F_1 score

		Modality						Average F_1
		TXT-IMG		TXT		IMG		
Query Type	T	2 runs	0.4621	14 runs	0.5905	1 run	0.0951	0.5462
	T-CT-CD-I	10 runs	0.5729	2 runs	0.4579	3 runs	0.1296	0.4689
	T-CT	2 runs	0.7214	13 runs	0.6071	-		0.6233
	T-CT-I	8 runs	*0.7344*	1 run	0.6842	-		*0.7288*
	T-CT-CD	2 runs	0.6315	7 runs	0.5688	-		0.5827
	I	4 runs	0.6778	1 run	0.6741	3 runs	*0.1786*	0.4901
	T-I	6 runs	0.7117	1 run	*0.7492*	-		0.7171
	CT-I	2 runs	0.6925	-		-		0.6925
	CT	-		2 runs	0.6687	-		0.6687

It is interesting to note that the highest F_1 score was different for each modality. A combination of T-CT-I had the highest score in TXT-IMG modality. In the TXT modality, a combination of T-I scored the highest, with T-CT-I following on the second place. However, since only one run used the T-I, it was not enough to provide a conclusion about the best run. Calculating the average F_1 score regardless of diversity shows that the best runs are achieved using a combination of Title, Cluster Title and Image. Using all tags in the queries resulted in the worst performance.

5 Conclusions

This paper has reported the ImageCLEF Photo Retrieval Task for 2009. Still focusing on the topic of diversity, this year's task introduced new challenges to the participants, mainly through the use of a much larger collection of images than used in previous years and by other tasks. Queries were released as two 'types': the first type of queries included information about the kind of diversity expected in the results; the second type of queries not providing this level of detail.

The number of registering participants in this year was the highest of all the ImageCLEFPhoto tasks since 2003. Nineteen participants submitted a total of 84 runs, which were then categorised based on the query fields used to find information, and the modalities being used. The result showed that participants were able to present a diverse result without sacrificing precision. In addition, results showed the following:

- Information about the cluster title is essential for providing diverse results, as this enables participants to correctly present images based on each cluster. When the cluster information was not being used, the cluster recall score is proven to drop, which showed that participants need better approach to predict the diversity need in it.

- A combination of Title, Cluster Title and Image was proven to maximize the diversity and relevance of the search engine.

- Using mixed modality (text and image) in the runs managed to achieve the highest F_1 compared to using only text or image features alone.

Considering the increasing interest of participants in ImageCLEFPhoto, the creation of the new collection was seen as a big achievement in providing a more realistic framework for the analysis of diversity and evaluation of retrieval systems aimed at promoting diverse results. The findings from this new collection were found to be promising and we plan to make use of other diversity algorithms in the future to enable evaluation to be done more thoroughly.

Acknowledgments

We would like to thank Belga Press Agency for providing us the collection and query logs and Theodora Tsikrika for the preprocessed queries which we used as the basis for this research.

The work reported has been partially supported by the TrebleCLEF Coordination Action, within FP7 of the European Commission, Theme ICT-1-4-1 Digital Libraries and Technology Enhanced Learning (Contract 215231).

References

[1] Tsikrika, T. 2009. Queries Submitted by Belga Users in 2008.

[2] Paramita, M. L, Sanderson, M., and Clough, P. 2009. Developing a Test Collection to Support Diversity Analysis. SIGIR 2009 Workshop: Redundancy, Diversity, and Interdependent Document Relevance, July 23rd, Boston, Massachusetts, USA.

[3] Arni, T., Clough, P., Sanderson, M., and Grubinger, M. 2008. Overview of the ImageCLEFPhoto 2008 Photographic Retrieval Task. Cross Language Evaluation Forum.