UNED at iCLEF 2009: Analysis of Multilingual Image Search Sessions

Víctor Peinado, Fernando López-Ostenero and Julio Gonzalo
NLP & IR Group, ETSI Informática, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
{victor, flopez, julio}@lsi.uned.es

Abstract

In this paper we summarize the analysis performed on the logs of multilingual image search provided by iCLEF09 and its comparison with the logs released in the iCLEF08 campaign. We have processed more than one million log lines in order to identify and characterize 5,243 individual search sessions.

We focus on the analysis of users' behavior and their performance trying to find possible correlations between: a) the language skills of the users and the annotation language of the target images; and b) the final outcome of the search session.

We have observed that the proposed task can be considered as easy, even though users with no competence in the annotation language of the images tend to perform more interactions and to use cross-language facilities more frequently. Usage of relevance feedback is remarkably low, but successful users use it more often.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries; H.5.2 User Interfaces

General Terms

cross-language information retrieval, interactive systems, image search, known-item retrieval task

Keywords

flickr, images, log analysis

1 Introduction

In this paper we summarize the analysis performed on the logs of multilingual image search provided in the iCLEF 2009 track [2] and its comparison with the logs released in the iCLEF 2008 campaign [1].

In the search logs provided by the organizers individual search sessions can be easily identified. Each session starts when a registered user is shown a target image and finishes when the user finds the image or decides to give up. The logs collects all the interactions occurred in the meantime: monolingual and multilingual queries launched, query refinements, navigation across the results ranking, hints showed by the system, usage of the personal dictionaries and other cross-language facilities, etc. These logs are automatically generated by the FlickLing search engine. Please see [3] for a complete description on the interface's functionalities and the logs.

Last year [5] we focused on the analysis of possible correlations between the language skills of the users and the annotation language of the target images, along with the usage of some of the specific cross-language facilities FlickLing features. In this work we are going to focus on the analysis of users' behavior and their performance trying to find possible correlations between: a) the language skills of the users and the annotation language of the target images; and b) the final outcome of the search session. Being aware of the differences between both groups of users involved in an interactive experiment and between both pools of images used, we are replicating the analysis trying to find out new correlations and reinforce or discard the evidences observed.

The remainder of the paper is as follows: Section 2 describes the processing tasks and the characterization of the search sessions performed on the iCLEF logs. Next, we discuss some correlations found between our users' search behavior and their profile according to their language skills (Sections 3) and the final outcome of their search sessions (Section 4). Then, in Section 5 we present additional data of our study based on the questionnaires collected during the experiments. Finally, in Section 6 we draw some general conclusions and move forward to propose future work lines.

2 iCLEF Logs Processing

The logs provided by the iCLEF organization in 2009 were considerably smaller than last year's corpus. Table 1 shows some of the most relevant statistics of both raw logs.

	2008	**2009**
registered users	305	130
log lines	$1,483,806$	$617,947$
valid search sessions	$5,101$	$2,410$
images found	$4,033$	$2,149$
images not found	$1,068$	261
hints asked	$11,044$	$5,805$
monolingual queries	$37,125$	$13,037$
multilingual queries	$36,504$	$17,872$
promoted translations	584	725
penalized translations	215	353
image descriptions shown	418	100

Table 1: Statistics of the logs provided by the iCLEF organization

Of course, we had many users who registered and tried just a few searches. As last year, for the current analysis, we are going to focus only on those users who, regardless of their final outcome, were able to complete at least 15 search sessions and filled in the overall questionnaire. We think that these users finished the experiment (even though they could go on searching at will) and became experienced enough using FlickLing. Table 2 shows some of the most relevant statistics of both logs considering only the mentioned sub-sets of users.

Notice that we are going to analyze more than one million log lines generated by 98 users and containing $5,243$ search sessions and more than $62,000$ queries. Comparing a collection of logs generated in an interactive image search experiment with different users and two different sets of target images is not straightforward, but we think these figures are large enough to reach quantitatively meaningful conclusions.

So, we have processed the logs in order to obtain a rich characterization of the search sessions: the user and her behavior, the target image and the usefulness of the search and translations facilities provided by flickling. As in our previous work [4], we have extracted 115 features for each session, capturing the complete user's profile according to her language skills, the target image's profile, and the usage of the interface's functionalities. The first hint provided by the system when the user decides to quit is always the language in which the target image is annotated. Since

	2008	**2009**
considered users	65	33
log lines	841, 957	357, 703
valid search sessions	3, 640	1, 603
images found	2, 983	1, 439
images not found	657	164
hints asked	8, 093	3, 886
monolingual queries	23, 060	8, 461
multilingual queries	20, 607	10, 463
promoted translations	223	525
penalized translations	70	246
image descriptions shown	126	42

Table 2: Statistics of the sub-sets of logs analyzed

this fact may turn the initial fully multilingual search (target images can be tagged in up to six different languages, e.g.: Dutch, English, French, German, Italian and Spanish) into a bilingual or monolingual search (depending on the user's language skills), we have also tracked the user's behavior before and after asking for this first hint.

In the following sections we present the analyses performed on these two sub-sets of search sessions according to the language skills of the users (Section 3) and considering the final outcome of the search sessions (Section 4).

3 Analysis According to Language Skills

We have divided our search sessions into three different 'profiles' according to the user's language skills with respect to the annotation language of the target image. On one hand, "active" denotes the sessions where the image was annotated in a language in which the user was able to read and write fluently. On the other hand, "passive" sessions are those where the target language was partially understandable by the user, but the user could not make queries in that language (e.g. images annotated in Italian for most Spanish or French speakers). Finally "unknown" refers to sessions when the image is annotated in languages completely unfamiliar for the user.

3.1 Users' Behavior

While the iCLEF08 corpus has samples enough under these categories (2, 345 sessions for active, 535 for passive and 760 for unknown), iCLEF09 corpus has no active sessions at all and has a great majority of unknown sessions (only 18 are passive and 1585 are unknown). The explanation for this is in the different characteristics of the target images proposed each year. Last year the image corpus was fully multilingual but most of the images could be easily found by simply searching in English and Spanish, the most popular languages among our users. This year, on the contrary, the image corpus was collected trying to avoid images annotated in English and stressing carefully on Dutch and German. Our users were coming basically from Romania, Italy and Spain, with little knowledge in these languages.

Table 3 shows the number of samples per profile, the average values for success rate (was the image found?) and the average number of hints requested per search session for each year's logs, along with the aggregate values.

According to the figures, it seems that the degree of success was high in all cases. In the iCLEF08 corpus, active and passive speakers performed similarly (passive users asking for more hints, though): they successfully found the target image 84% and 82% of the times, respectively. On the other hand, as expected, users with no competence in the annotation language obtained 73% of success rate and asked for more hints (2.42).

iCLEF08			
result	**samples**	**success rate**	**# hints requested**
active	2,345	85%	2.14
passive	535	82%	2.22
unknown	760	73%	2.42
iCLEF09			
result	**samples**	**success rate**	**# hints requested**
active	0	-	-
passive	18	78%	1.22
unknown	1,585	90%	2.43
iCLEF08 + iCLEF09			
result	**samples**	**success rate**	**# hints requested**
active	2,345	85%	2.14
passive	553	82%	2.12
unknown	2,345	84%	2.45

Table 3: User's behavior according to language skills: average success rate and hints requested

In the iCLEF09 corpus, the division in profiles does not allow to find clear correlations because of the lack of samples. Unknown users, nonetheless, were able to successfully find the image 90% of the times, while asking for 2.43 hints, a smiliar figure compared to iCLEF08. It's worth noticing that hints in iCLEF09 were more specific and concrete than in iCLEF08. Thus, even though most of the target images were annotated in an unknown language, asking for hints was definitely more useful this year.

Finally, in the the aggregate figures, it can be observed that while all three profiles present a similar success rate, the unknown users asks for more hints than users with some degree of competence in the annotation language of the image.

3.2 Cognitive Effort

We have grouped under the name "cognitive effort" some of the most usual interactions occurred in a traditional search interface, namely: launching queries, exploring the ranking of results beyond the first page (each page contains 20 items), and using relevance feedback (words provided by Flickr related to the query terms, and the tags associated to each image retrieved in the ranking of results). So, Table 4 shows the figures related to these interactions for each user profile in both FlickLing's monolingual and multilingual environments.

In the iCLEF08 logs, as expected, active and passive users launch more queries in the monolingual environment, while unknown users, who are supposed to need some translation functionalities to find the image, launch more multilingual queries using FlickLing's facilities. As far as the ranking exploration is concerned, the same pattern appears: active and passive users cover more ranking pages while querying in monolingual and unknown users explore the ranking more deeply while querying in multilingual.

Analyzing the iCLEF09 results, we cannot draw any clear conclusions but if we ignore the 18 samples corresponding to passive users, we find the unknown users again performed more interactions in the multilingual environment: more queries launched and more ranking explorations.

Usage of relevance feedback facilities, as shown in previous works (see [5]), is very low for both logs collections. But even with small variations, active and passive players used relevance feedback more often with monolingual searches and unknown players used it more often in the multilingual environment.

Analyzing the aggregate data we can maintain the following conclusions: active and passive users employed more cognitive effort in monolingual searches while unknown users needed more cognitive effort in multilingual searches in order to reach a similar performance, as shown in Section 3.1.

iCLEF08						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
active	4.03	3.28	2.09	1.92	0.03	0.03
passive	4.16	3.31	2.83	2.24	0.05	0.02
unknown	3.81	4.02	2.36	2.81	0.07	0.09
iCLEF09						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
active	-	-	-	-	-	-
passive	4.72	11.06	2.78	11.11	0	0
unknown	3.48	3.89	1.76	2.43	0.01	0.03
iCLEF08 + iCLEF09						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
active	4.03	3.28	2.09	1.92	0.03	0.03
passive	4.18	3.56	2.83	2.53	0.05	0.02
unknown	3.57	3.91	1.96	2.55	0.03	0.05

Table 4: Cognitive effort according to language skills: typed queries, ranking exploration and usage of relevance feedback

3.3 Usage of Specific Cross-Language Refinement Facilities

The dictionaries used by FlickLing were not optimal. In order to cover the six languages considered in the experiment, freely-available general-purpose dictionaries were used. To rectify some of the translation errors, FlickLing allows users to promote and penalize the translations appearing in the general dictionaries of its multilingual environment. This changes are incorporated into a personal dictionary for each user and do not affect other players' translations. When characterizing the search sessions, we also took into consideration the usage of this functionality by our users.

In general, the usage of the personal dictionary was low. Table 5 shows the average percentage of search sessions in which users manipulated their personal dictionary by adding new translations, promoting good translation options and removing bad ones, and the average query terms modified by these manipulations.

iCLEF08		
competence	**dictionary manipulations**	**query terms modified**
active	0.06	0.04
passive	0.05	0.03
unknown	0.17	0.11
iCLEF09		
competence	**dictionary manipulations**	**query terms modified**
active	-	-
passive	6.56	1.67
unknown	0.4	0.16
iCLEF08 + iCLEF09		
competence	**dictionary manipulations**	**query terms modified**
dictionary manipulations	**query terms modified**	
active	0.06	0.04
passive	0.27	0.08
unknown	0.33	0.14

Table 5: Usage of specific cross-language refinement facilities according to language skills

In iCLEF08, unknown users manipulated their personal dictionary about three times (0.17) more often than active (0.06) and passive (0.05) players, and consequently the number of query terms modified was also higher (0.11). If we compare both log collections, we observe how in iCLEF09, where the usage of cross-language facilities was more expected, was also increased (0.4). As far as the aggregate data are concerned, we can observe that the more lack of language skills a user has, the more she uses these cross-language facilities.

4 Analysis According to Search Session's Outcome

In the following sections we are going to analyze users' behavior according to the final outcome of the search sessions. In order to find some correlations about the most successful strategies used by our users, we are going to divide the sessions into two categories: on one hand, "success" refers to those sessions where users were, with or without hints, able to find the proposed target image. On the other hand, "fail" refers to those sessions where the user decided to quit before finding the image.

4.1 Users' Behavior

As we saw in Section 3.1, we are going to analyze users' behavior but stressing now on the final outcome of the search sessions. If we see Table 6, the first detail to be noted is the number and percentage of samples of each category: 81.95% of success samples in iCLEF08, 89.77% in iCLEF09 and 84.34% in the aggregate results confirm that finding the proposed images was an easy task.

iCLEF08			
result	**samples**	**%**	**# hints requested**
success	2,983	81.95%	2.32
fail	657	18.05%	1.74
iCLEF09			
result	**samples**	**%**	**# hints requested**
success	1,439	89.77%	2.38
fail	164	10.23%	2.77
iCLEF08 + iCLEF09			
result	**samples**	**%**	**# hints requested**
success	4,422	84.34%	2.34
fail	821	15.66%	1.95

Table 6: User's behavior according to search session outcome: average success rate and hints requested

Regarding the average number of hints requested, users in successful sessions asked for 2.32 and 2.38 hints in iCLEF08 and iCLEF09, respectively. Users in failed sessions asked for a similar quantity of hints in iCLEF09 (2.77), while in iCLEF08 the number of hints is lower (1.74).

Finally, in the aggregate results we can observe that asking for hints seems to have been a good strategy to find the target image, in spite of the score loss, since users in successful sessions asked for 2.34 hints compared to 1.95 for failed sessions.

4.2 Cognitive Effort

Analyzing the cognitive effort with respect to the outcome of the search session, our aim is to find some correlation about what strategy was the most convenient for our users to find the images in the iCLEF experiment proposed.

As shown in Table 7, in the iCLEF08 logs, successful users launched more queries in the monolingual environment than in the multilingual one (4.05 vs. 3.36), while unsuccessful players does not show differences 3.76 vs. 3.79). On the other hand, in the iCLEF09 logs, successful users launched more multilingual queries than monolingual (4.02 vs. 3.65). This can be explained, as mentioned above, because of the kind of the image collection, which was designed to force the multilingual searches. This fact can also be seen in the number of explorations of the ranking, slightly higher than in iCLEF09 (2.48 and 2.93 vs. 2.13 and 2.26). Lastly, in general, users in failed sessions seems to have performed more interactions in the monolingual environment.

iCLEF08						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
success	4.05	3.36	2.22	2.13	0.05	0.04
fail	3.76	3.79	2.39	2.26	0.05	0.02
iCLEF09						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
success	3.65	4.02	1.89	2.48	0.02	0.03
fail	1.96	3.23	0.79	2.93	0.01	0.02
iCLEF08 + iCLEF09						
competence	**typed queries**		**ranking exploration**		**relevance feedback**	
	mono	**multi**	**mono**	**multi**	**mono**	**multi**
success	3.92	3.58	2.11	2.24	0.04	0.04
fail	3.4	3.67	2.07	2.4	0.04	0.02

Table 7: Cognitive effort according to the search session outcome: typed queries, ranking exploration and usage of relevance feedback

As the last columns of the table show, the usage of relevance feedback was very low in both categories, being higher in monolingual in iCLEF08 and multilingual in iCLEF09. In general, but still with little differences, successful users tended to use relevance feedback more frequently.

4.3 Usage of Specific Cross-Language Refinement Facilities

Finally, regarding the manipulation of the personal dictionary (see Table 8), successful users in iCLEF08 used it slightly more often than those who failed (0.08 vs. 0.06). In iCLEF09, the general usage is much more higher, but the pattern is reproduced upside down: unsuccessful players tended to manipulate their dictionaries more often (0.62 vs. 0.46).

In the aggregate data corresponding to both logs we can observe that successful players used this functionality more frequently (0.2 vs. 0.17).

5 Questionnaire analysis: Users' Perception on the Task

Along with the interactions of the users and the information of the search sessions, iCLEF logs also contain two types of questionnaires: one is shown every time the user finishes a search session and it contains questions about the target image and the development of the search. The other questionnaire is shown when the user has completed 15 search sessions and raises overall questions about the task itself, the usefulness of the interface functionalities and the user's performance.

In this analysis we are going to focus only on the former one, specially in the following questions:

Which, in your opinion, are the most challenging aspects of the task? 83% of participants from iCLEF08 and 85% from iCLEF09 agree or strongly agree that "Selecting/finding appropriate translations for the terms in my query" was the most challenging aspect of the task.

iCLEF08		
competence	**dictionary manipulations**	**query terms modified**
success	0.08	0.05
fail	0.06	0.05
iCLEF09		
competence	**dictionary manipulations**	**query terms modified**
success	0.46	0.17
fail	0.62	0.18
iCLEF08 + iCLEF09		
competence	**dictionary manipulations**	**query terms modified**
dictionary manipulations	**query terms modified**	
success	0.2	0.09
fail	0.17	0.07

Table 8: Usage of specific cross-language refinement facilities according to the search session outcome

Users from both years also agree or strongly agree with other answers such as "Finding the correct terms to express an image in my own native language", "Handling multiple target languages at the same time" and "Finding the target image in very large sets of results" in about 80% of the cases.

Which interface facilities did you find most useful? In both years, cross-language functionalities such as the automatic translation and the possibility of maintaining a personal dictionary are more valued than relevance feedback facilities, specially among the iCLEF09 users. 80% of the users from 2009 agree with the usefulness of the personal dictionary, against the 59% who agree with the usefulness of the additional query terms suggested by the system.

Which interface facilities did you miss? Up to seven different facilities not implemented in the current version of FlickLing are proposed in this question. Among the iCLEF08 users, the most popular answers with an agreement rate about 75% were "The classification of search results in different tabs according to the image caption languages" and "A system able to select the translations for my query terms better".

As far as the iCLEF09 users are concerned, the most popular answers with more than 80% of support were, along with "A system able to select the translations for my query terms better", "Bilingual dictionaries with a better coverage" and "Detection and translation of multi-word expressions". These answers seem to be accordance with the fact that iCLEF09 users needed to interact more frequently in a multilingual environment with cross-language tools that could be improved.

How did you select/find the best translations for your query terms? Again a question in accordance to the different users' and images' profiles in both campaigns. While the most popular answer for the iCLEF08 users was "Using my knowledge of target languages whenever possible" (around 90%), iCLEF09 users opted for "Using additional dictionaries and other on-line sources" in 82% of the cases. .

6 Conclusions and Future Work

In this paper we have summarized the analysis performed on the logs of multilingual image search provided by iCLEF09 and its comparison with the logs released in the iCLEF08 campaign. We have processed more than one million log lines in order to identify and characterize $5,243$ individual search sessions. Each session starts when a registered user is shown a target image and finishes when the user finds the image or decides to give up. Besides, the logs collects all the interactions occurred in the meantime: monolingual and multilingual queries launched, query refinements, navigation across the results ranking, hints showed by the system, usage of the personal dictionaries and other cross-language facilities, etc.

In this work we have focused on the analysis of users' behavior and their performance trying to find possible correlations between: a) the language skills of the users and the annotation language of the target images; and b) the final outcome of the search session.

Among the conclusions observed in this work, we can mention:

- The proposed task was easy, since all users' profiles reach more than 80% of success rate. Users with no competence in the annotation language of the image tend to ask for more hints.
- Users with some knowledge in the annotation language of the images employ more cognitive effort in monolingual searches, while users without skills need more cognitive effort in multilingual searches in order to reach a similar performance.
- As expected, the more lack of language skills a user has, the more she uses cross-language facilities.
- Given the features of the two images collections, in iCLEF08, where most of the images were annotated in known languages, successful users launched more queries in the monolingual environment. On the other hand, in iCLEF09, where multilingual needs were forced on purpose, successful users launched more multilingual queries.
- Usage of relevance feedback is remarkably low, but successful users tended to use it more frequently.
- Questionnaires show that cross-language facilities are seen as very positive when proposed in a multilingual search scenario.
- The answers collected in the questionnaires are in accordance with the fact that iCLEF09 users needed to interact more frequently in a multilingual environment with cross-language tools that could be improved.

As part of the future work, we're currently widening this study analyzing users' behavior across time as they go ahead in the experiment finishing more and more search sessions, in order to find useful correlations about how they learn to interact with the system and how they test different search strategies.

Acknowledgements

This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and the Spanish Government under project Text-Mess (TIN2006-15265-C06-02).

We would also like to thank Javier Artiles for his intensive work during the implementation of the FlickLing interface and all the collaborators involved in collecting the image corpus and the testing stage.

References

[1] Gonzalo, J., Clough, P., Karlgren, J.: Overview of iCLEF 2008: search log analysis for Multilingual Image Retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008), Aarhus, Denmark. LNCS vol. 5706, Springer Verlag. 2009.

[2] Gonzalo, J., Clough, P., Karlgren, J.: Overview of iCLEF 2009. This volume.

[3] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F.: FlickLing: a multilingual search interface for Flickr. In: Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.

[4] Peinado, V., Gonzalo, J., Artiles, J., López-Ostenero, F.: UNED at iCLEF 2008: Analysis of a Large Log of Multilingual Image Searches in Flickr. In: Working Notes for the CLEF 2008 Workshop. Aarhus, Denmark, September 17-18. 2008.

[5] Peinado, V., Gonzalo, J., Artiles, J., López-Ostenero, F.: Log Analysis of Multilingual Image Search in Flickr. In: Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008), Aarhus, Denmark. LNCS vol. 5706, pp. 236-242. Springer Verlag. 2009.