Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy environment

Julio Gonzalo Víctor Peinado
UNED
Spain
{julio,victor}@lsi.uned.es

Paul Clough
U. Sheffield
United Kingdom
p.d.clough@sheffield.ac.uk

Jussi Karlgren
SICS
Sweden
jussi@sics.se

Abstract

This paper summarises activities from the iCLEF 2009 task. As in 2008, the task was organised based on users participating in an interactive cross-language image search experiment. Organizers provided a default multilingual search system (Flickling) which accessed images from Flickr, with the whole iCLEF experiment run as an online game. Interaction by users with the system was recorded in log files which were shared with participants for further analyses, and provide a future resource for studying various effects on user-orientated cross-language search. In total six groups participated in iCLEF with different approaches, ranging from pure log analysis to specific experiment designs using the Flickling interface.

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.4 [**Information Systems Applications**]: H.4.m Miscellaneous

General Terms

interactive information retrieval, cross-language information retrieval

Keywords

iCLEF, Flickr, log analysis, multilingual image search, user studies, multilingual tag search

1 Introduction

iCLEF is the interactive track of CLEF (Cross-Language Evaluation Forum), an annual evaluation exercise for Multilingual Information Access systems. In iCLEF, Cross-Language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language.

Since 2006, iCLEF has moved away from news collections (a standard for text retrieval experiments) in order to explore user behaviour in scenarios where the necessity for cross-language search arises more naturally for the average user. We chose Flickr, a large-scale, web-based image database based on a large social network of WWW users sharing over two billion images, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments.

Over the last years, iCLEF participants have typically designed one or more cross-language search interfaces for tasks such as document retrieval, question answering or text-based image retrieval. Experiments were hypothesis-driven, and interfaces were studied and compared using controlled user populations under laboratory conditions. This experimental setting has provided valuable research insights into the problem, but has a major limitation: user populations are necessarily small in size, and the cost of training users, scheduling and monitoring search sessions is very high. In addition, the target notion of relevance does not cover all aspects that make an interative search session successful; other factors include user satisfaction with the results and usability of the interface.

The main novelty of the iCLEF 2008 shared experience, which has been kept in 2009, was to focus on the shared analysis of a large search log from a single search interface provided by the iCLEF organizers. The focus is, therefore, on search log analysis rather than on system design. The idea is to study the behaviour of users in an (almost) naturalistic search scenario, having a much larger data set than in previous iCLEF campaigns. The search interface provided by iCLEF organizers is a basic cross-language retrieval system to access images in Flickr, presented as an online game: the user is given an image, and she must find it again without any a-priori knowledge of the language(s) in which the image is annotated. Game-like features are intented to engage casual users and therefore increase the chances of achieving a large, representative search log.

The iCLEF 2009 task is the same as in 2008, the only difference being the approach to select the target images (the topics for our task). In 2008 a large log was harvested, but in over half of the search sessions the user had active language skills in the target language, and the situations were the user has only passive or no abilities in the target language were underepresented. The reason was that many images in the target set had annotations in English (plus other languages in many cases), and the set of users (over 200 active searchers) tend to have English as a native or at least as a well-known language. Therefore, this year we explicitly avoided images annotated in English to increase the chances of having search sessions in unknown languages.

The structure of the rest of the paper is as follows: Section 2 describes the task guidelines (and can be skipped by readers familiarized with the iCLEF 2008 task); Section 3 describes the features of the search log distributed to participants. In Section 4 we summarize the participation in the track and give some conclusions about the experience.

2 Task guidelines

The task is exactly the same as in 2008, and the differences lie in the search log collected (target images, set of registered users, etc.) Readers which are familiarized with iCLEF 2008 can safely skip this Section.

2.1 Search task definition

First of all, the decision to use Flickr as the target collection is based on (i) the inherent multilingual nature of the database, provided by tagging and commenting features utilised by a worldwide network of users, (ii) although it is in constant evolution, which may affect reproducibility of results, the Flickr search API allows the specification of timeframes (e.g. search in images uploaded between 2004 and 2007), which permits defining a more stable dataset for experiments; and (iii) the Flickr search API provides a stable service which supports full boolean queries, something which is essential to perform cross-language searches without direct access to the index.

For 2008, our primary goal was harvesting a large search log of users performing multilingual searches on the Flickr database. Rather than recruiting users (which inevitably leads to small populations), we wanted to publicize the task and attract as many users as possible from all around the world, and engage them with search. To reach this goal, we needed to observe some restrictions:

- The search task should be clear and simple, requiring no a-priori training or reading for the casual user.

- The search task should be engaging and addictive. Making it an online game - with a rank of users - helps achieve that, with the rank providing a clear indication of success.
- There should be no need for manual judgements in order to establish the success of a search session, in order to avoid discouraging delays in the online game rankings.
- It should have an adaptive level of difficulty to prevent novice users from being discouraged, and to prevent advanced users from being unchallenged.
- The task should be naturally multilingual.

We decided to adopt a known-item retrieval search task: the user is given a raw (unnanotated) image and the goal is to find the image again in the Flickr database, using a multilingual search interface provided by iCLEF organizers. The user does not know in advance in which languages the image is annotated; therefore searching in multiple languages is essential to get optimal results. Although the task is probably not the most natural one (thematic-based searches are probably more common than "stuff I've seen before" search needs), it has the definitive advantage of not requiring manual judgements, and that makes possible to keep an instantly updated user ranking.

Indeed the task is organized as an online game: the more images found, the higher a user is ranked. In case of ties, the ranking will also depend on precision (number of images found / number of images attempted). At any time the user can see the "Hall of Fame" with a rank of all registered users.

Depending on the image, the source and target languages, this can be a very challenging task. To have an adaptive level of difficulty, we implemented a hints mechanism. At any time whilst searching, the user is allowed to quit the search (skip to next image) or ask for a hint. The first hint is always the target language (and therefore the search becomes mono or bilingual as opposed to multilingual). The rest of the hints are keywords used to annotate the image. Each image found scores 25 points, but for every hint requested, there is a penalty of 5 points. The hint mechanism proved essential to engage users in 2008 and even more in 2009 (for reasons explained later).

Initially a five minute time limit per image was considered, but initial testing indicated that such a limitation was not natural and had a deep impact on users' search behaviour. Therefore we decided to remove time restrictions from the task definition.

2.2 Search interface

We designed the so-called *Flickling* interface to provide a basic cross-language search front-end to Flickr. Flickling is described in detail in [1]; here we will summarize its basic functionalities:

- User registration, which records the user's native language and language skills in each of the six European languages considered (EN, ES, IT, DE, NL, FR).
- Localization of the interface in all six languages.
- Two search modes: mono and multilingual. The latter takes the query in one language and returns search results in up to six languages, by launching a full boolean query to the Flickr search API.
- Cross-language search is performed via term-to-term translations between six languages using free dictionaries (taken from: `http://xdxf.revdanica.com/down`).
- A term-to-term automatic translation facility which selects the best target translations according to (i) string similarity between the source and target words; (ii) presence of the candidate translation in the suggested terms offered by Flickr for the whole query; and (iii) user translation preferences.

- A query translation assistant that allows users to pick/remove translations, and add their own translations (which go into a personal dictionary). We did not provide back-translations to support this process, in order to study correlations between target language abilities (active, passive, none) and selection of translations.
- A query refinement assistant that allows users to refine or modify their query with terms suggested by Flickr and terms extracted from the image rank. When the term is in a foreign language, the assistant tries to display translations into the user's preferred language to facilitate feedback.
- Control of the game-like features of the task: user registration and user profiles, groups, ordering of images, recording of session logs and access to the hall of fame.
- Post-search questionnaires (launched after each image is found or failed) and final questionnaires (launched after the user has searched fifteen images, not necessarily at the end of the experience).

2.3 Participation in the track

As in 2008, iCLEF 2009 participants can essentially adopt two types of methodology: (1) analyse log files based on all participating users (which is the default option) and, (2) perform their own interactive experiments with the interface provided by the organizers. CLEF individuals registered in the interface as part of a team, so that a ranking of teams is produced in addition to a ranking of individuals.

2.3.1 Generation of search logs

Participants can mine data from the search session logs, for example looking for differences in search behaviour according to language skills, correlations between search success and search strategies, etc.

2.3.2 Interactive experiments

Participants can recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc. iCLEF organizers provided assistance with defining appropriate user groups and image lists, for example, within the common search interface. Besides these two options, and given the community spirit of iCLEF, we were open to groups having their own plans (e.g. testing their own interface designs or using a specific set of images) as long as they did not change the overall shared search task (known-item search on Flickr).

3 Dataset: Flickling search logs

Search logs were harvested from the Flickling search interface between May and June 2009 (see [1] for details on the logs content and syntax). In order to entice a large set of users, the "CLEF Flickr Challenge" was publicized in Information Access forums (e.g. the SIG-IR and CLEF lists), Flickr blogs and general photographic blogs. As in 2008, we made a special effort to engage the CLEF community in the experience, with the goal of getting researchers closer to the CLIR problem from a user's perspective. To achieve this goal, CLEF organizers agreed to award two prizes consisting of free registrations for the workshop: one for the best individual searcher and one for the best scoring CLEF group.

Overall, 130 users registered for the task, for a total of 2527 search sessions, many of them ending in success (2149). There were 19 native languages in our user set, with this distribution:

46 Spanish, 38 Romanian, 10 English, 9 Italian, 4 Persian/Farsi, 4 German, 3 Chinese, 2 Finnish, 2 Catalan, 2 Basque, 2 Arabic, 1 Danish, 1 Vietnamese, 1 Malay, 1 Russian, 1 Greek and 1 Belarusian.

Apart from general users, the group affiliation revealled two dominant user profiles: university researchers and students (most of them in Computer Science) and photography fans.

The 2008 search log was skewed towards "active" search sessions (where users had active skills in some of the languages used to annotate the image). Therefore this year we changed the methodology to select the target images, excluding those which had annotations in English, and reducing the number of images annotated in Spanish (because it was a well represented native language in our user base). The strategy was too successful: we harvested 1585 search sessions where the target language was unknown to the user, 18 where the user had passive abilities (i.e. could read results but not write queries), and none where the user had active skills in the target language. That makes this search log an excellent tool to study the behaviour of users searching in foreign language, but it can hardly be used to compare the three profiles. We also found that the combination of users and images is so different from the 2008 experience that merging the two search logs, even if the task is the same, is not advisable.

Overall, it has been possible to collect a large controlled multilingual search log, which includes both search behaviour (interactions with the system) and users' subjective impressions of the system (via questionnaires). This offers a rich source of information for helping to understand multilingual search characteristics from a user's perspective.

4 Participation and findings

Six sites submitted results for this year's interactive track: two newcomers (University of North Texas and Alexadru Ioan Cuza University, UAIC, in Romania) and four groups with previous experience in iCLEF: Universidad Nacional de Educación a Distancia (UNED), the Swedish Institute of Computer Science (SICS), Manchester Metropolitan University (MMU), and the University of Alicante.

University of Alicante [5] investigated whether there is a correlation between lexical ambiguity in queries and search success and, if so, whether explicit Word Sense Disambiguation can potentially solve the problem. To do so, they mined data from the search log distributed by the iCLEF organization, and found that less ambiguous queries lead to better search results and coarse-grained Word Sense Disambiguation might be helpful in the process.

UAIC [2] tried to find correlations between different search parameters using a subset of the search log consisting of searchers performed by a set of 31 users recruited fro the task (which were very active, performing almost 46% of all queries in the general search log). They did not find a clear connection between the results of over-achieving users and their particular actions, and they found hints of a possible (light) collaboration between them, which eventually makes our search log less reliable than initially thought.

Manchester Metropolitan University [3] tried to demonstrate the value in focusing on user's trust and confidence in the exploration of seeking behaviour to reveal users' perception of the tasks involved when searching across languages. Instead of focusing on log analysis, MMU recruited their own set of 24 users selected a specific set of three images (in Dutch, German and Spanish) and performed a qualitative and quantitative analysis including questionnaires, observational study of the search sessions, retrospective thinking aloud and interviews. Among other things, they found that variations in perceptions of searching and approach to using translations which is unrelated to the amount or type of help or guidance given. They also found that, in general, users only think about languages after asking for the first hint (i.e. the target language), facing cross-linguality only when it is inevitable.

UNED [4] tried to establish differences between users with active/passive/no knowledge of the target language, including search success and cognitive effort, and compared the results using search logs from 2008 and 2009. Unfortunately the skewed distribution of language profiles in 2009 did not permit direct comparisons and made results from the merged logs unreliable. UNED then

	Success ("foundImg")	Give up ("giveUp")
	2149	261
Time to resolution (average)	1 420 s	412 s
Reformulations (average)	110	29
`search`	3.7	6.3
Scroll actions		
`search`	1.8	1.3

Table 1: Some quantitative results distinguishing successful query sequences from failed ones. (Logs from 2009.)

worked on estabilishing successful search strategies when searching in foreign, unknown language. They found that the usage of cross-language search assistance features has an impact on search success, and that such features are highly appreciated by users.

University of North Texas [6] aimed at understanding the challenges that users face when searching for images that have multilingual annotations, and how they cope with these challenges to find the information they need. Similarly to MMU, instead of using the search log this group recruited their own set of six north american students and studied their search behaviour and subjective impressions using questionnaires, training, interviews and observational analysis. They found that users have strong difficulties using flickr tags, particularly when doing cross-language search, and that their typical session requires two hints: the target language and a keyword.

SICS has continued to investigate methods for how to study confidence and satisfaction of users. In previous years' studies, results have been somewhat equivocal; this year, some preliminary studies of the number of reformulations versus success rate have been performed. The SICS team found that the length of query sequences which eventually were successful were longer, indicating persistence when a search appears to be in the right direction. The number of query reformulations also correlate well with success: successful query sequences are a result of active exploration of the query space. *But* for users who persist in working with monolingual searches (`search` calls), the SICS team found that queries, firstly tended to be vastly less often reformulated to begin with, and that the successful sequences were more parsimonious than the failed ones (conversely from the `clsearch` calls): instead the number of scroll actions were much more frequent. This would seem to indicate that if users are fairly confident of a well put query, they will persist by scrolling through result lists. The figures in Table 1 are all statistically significant by the Mann Whitney U rank sum test ($p > 0.95$).

5 Conclusions

iCLEF 2009 has continued to run a large-scale interactive experiment as an online game to generate log files for further study. A default multilingual information access system developed by the organizers was provided to participants to lower the cost of entry and generate search logs recording user's interaction with the system and qualitative feedback about the search tasks and system (through online questionnaires). In addition, two groups have decided to replace (or extend) log analysis by recruiting their own set of users and employ the usual methodology (training, questionnaires, interviews, retrospective thinking aloud, observational studies) on them.

The search logs generated by the iCLEF track in 2008 and 2009 together are a reusable resource for future user-orientated studies of cross-language search behaviour, and we hope to see new outcomes in the near future coming from in-depth analysis of our logs. Researchers interested in this resource might contact the iCLEF organization (see http://nlp.uned.es/iCLEF) for details.

Acknowledgements

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

References

[1] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F.: FlickLing: a multilingual search interface for Flickr. In CLEF 2008 Workshop Notes, 2008.

[2] Cristea, F., Alexa, V. and Iftene, A. UAIC at iCLEF 2009: Analysis of Logs of Multilingual Image Searches in Flickr. In CLEF 2009 Workshop Notes, 2009.

[3] Vassilakaki, E., Johnson, F., Hartley, R.J., Randall, D.: Users' Perceptions of Searching in Flickling. In CLEF 2009 Workshop Notes, 2009.

[4] Peinado, López-Ostenero, F. and Gonzalo, J.: UNED at iCLEF 2009: Analysis of Multilingual Image Search Sessions. In CLEF 2009 Workshop Notes, 2009.

[5] Navarro-Colorado, Borja, Puchol-Blasco, M., Terol, Rafael M., Vázquez, S. and Lloret, E.: Lexical Ambiguity in Cross-Language Image Retrieval: a Preliminary Analysis. CLEF 2009 worskhop notes, 2009.

[6] Ruiz, M. and Chin, P. Users' Image Seeking Behaviour in a Multilingual Tag Environment. In CLEF 2009 Workshop Notes, 2009.