Overview of iCLEF 2008: Search Log Analysis for Multilingual Image Retrieval

Julio Gonzalo	Paul Clough	Jussi Karlgren
UNED	U. Sheffield	SICS
Spain	United Kingdom	Sweden
julio@lsi.uned.es	p.d.clough@sheffield.ac.uk	jussi@sics.se

Abstract

This paper summarises activities from the iCLEF 2008 task. In an attempt to encourage greater participation in user-orientated experiments, a new task was organised based on users participating in an interactive cross-language image search experiment. Organizers provided a default multilingual search system 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, providing a combined effort in generating results for a shared experiment on user-orientated cross-language retrieval.

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

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 interactive 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 has been 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 intended to engage casual users and therefore increase the chances of achieving a large, representative search log. More information can be found in [8].

The structure of the rest of the paper is as follows: Section 2 describes the task guidelines; 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

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.
- 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.

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.

Initially a five minute time limit per image was considered, but initial testing indicated that such a limitation was not natural and changed 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.[1]
- 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).

[1]Thanks go to the CLEF groups at the U. of Amsterdam, U. of Hildesheim, ELDA and CNR for providing native translations of the interface texts.

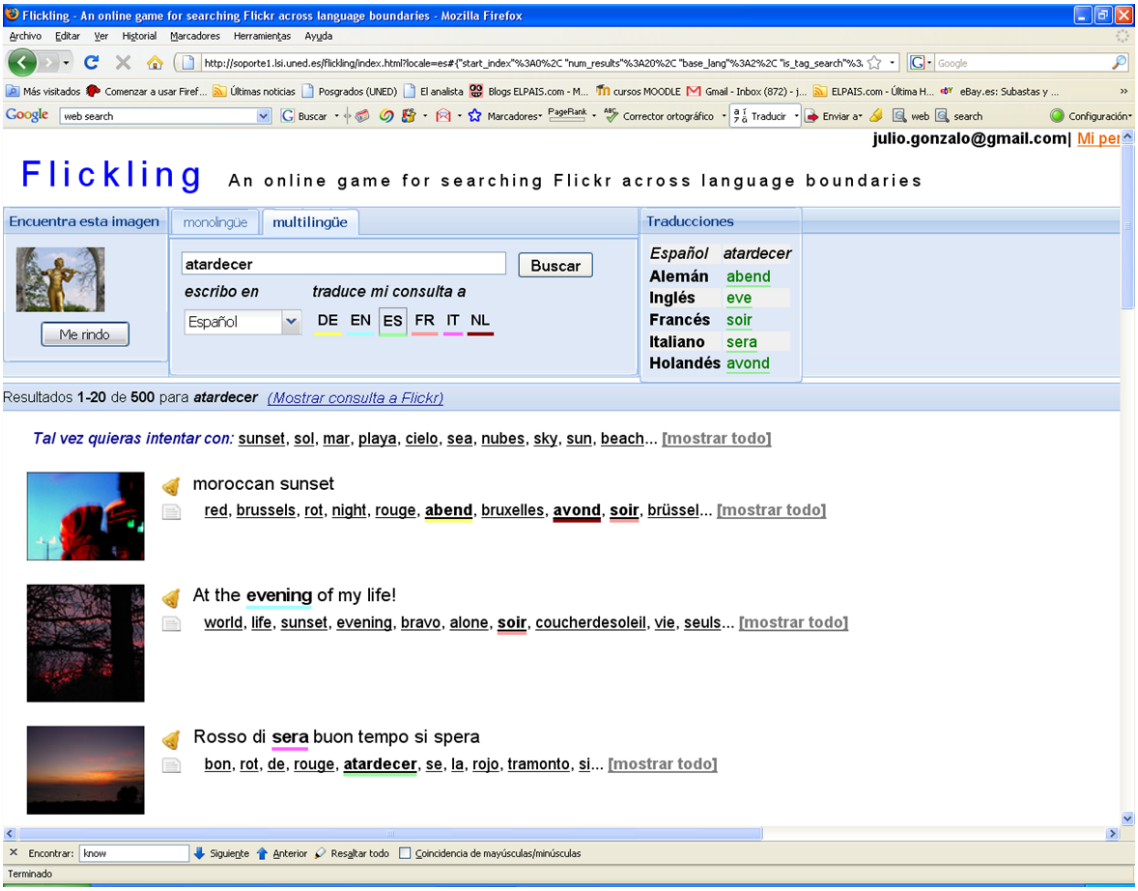

Figure 1: The Flickling search interface used to harvest search logs.

Figure 1 shows a snapshot of the search interface. Note that we did not intend to provide the best possible cross-language assistance to search the Flickr collection. As we wanted to focus on user behaviour - rather than on hypothesis testing for a particular interactive facility - our intention was to provide a standard, baseline interface that is not dependent on a particular approach to cross-language search assistance.

2.3 Participation in the track

Participants in iCLEF2008 can essentially do two tasks: (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 will register 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, or correlations between search success and search strategies.

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

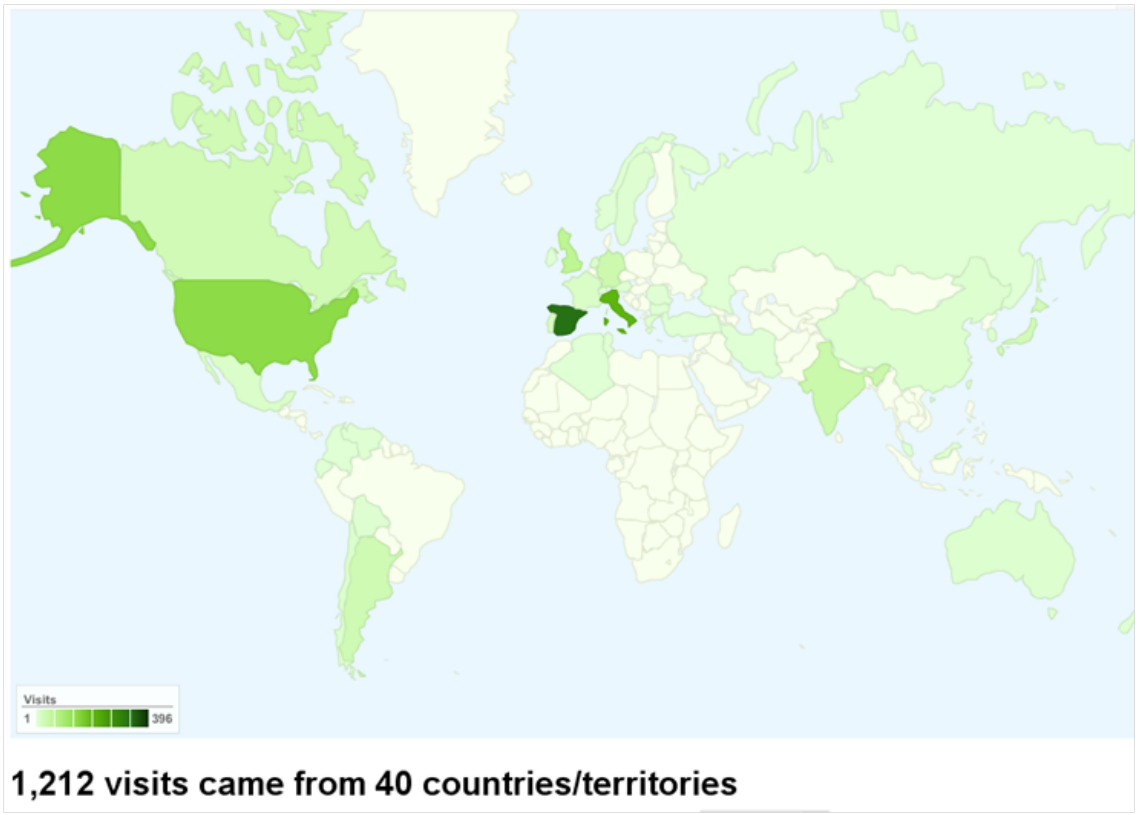

Figure 2: Geographic distribution of accesses in the search logs

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) 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 the beginning of May and the 15th of June 2008 (see [1] for details on the content and syntax of the logs). 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. 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 registration for the workshop: one for the best individual searcher and one for the best scoring CLEF group.

Dissemination was successful: during the log harvesting period, the interface was visited by useres from 40 different countries from Europe, the Americas, Asia and Oceania (see Figure 2). More than 300 people registered (around 230 were active searchers) and 104 performed searches for at least 10 different images. Out of them, 18 users attempted all 103 images considered for the task. Apart from general users, the group affiliation revealed at least three user profiles: researchers in Information Retrieval, linguistics students (most from the University of Padova) and photography fans (many entering from a Spanish blog specialized in photography, `dzoom.org.es`).

Profiles of user's language skills were very diverse, with a wide range of native and second language abilities. There was a total of 5101 complete search sessions (i.e. a user starts searching for an image and either finds the image or gives up), out of which the image was annotated in an active language (for the user) in 2809 cases, in an unknown language in 1566 cases, and in a passive language (when the user can partially read but cannot write) in 726 cases. Note that, even when the image is annotated in an active language for the user, this is not known by the user a-priori, and therefore the search behaviour is equally multilingual.

On average each search session included around four queries launched in the monolingual search mode, and four queries in the multilingual search mode. Overall, it was 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. A reusable data source has been produced for the first time since iCLEF first began.

4 Participation and findings

Six groups submitted results for this year's interactive track: Universidad Nacional de Educación a Distancia (UNED), the Swedish Institute of Computer Science (SICS), Manchester Metropolitan University (MMU), the University of Padua (UNIPD), University of Westminster, and the Indian Institute of Information Technology Hyderabad (IIIT-H). Studies ranged from exploring the effects of searcher background on results, studying how much attention searchers pay to language phenomena when searching images, how the effect of constraining the session might influence results, and examining logs to find evidence of user confidence in the search process.

UNED examined the effects of searcher competence in the target language and system learning effects, studying the logs and examining user responses to the questionnaires given to users at the completion of each completed or aborted task [5]. Analyses showed that when users had competence in the target language, their success at searching was higher; with passive knowledge user interaction showed similar success to those with active competence, but requiring more interactions with the system.

SICS studied the logs to find evidence of different levels of user confidence and competence in the behaviour exhibited and recorded in them [4]. The main conclusion is that to study these effects, the task design must be formulated to better capture and distinguish the difference between user decisions to terminate or continue a search.

MMU studied how users considered language and cross-linguistic issues during a session and how they switched between the cross-lingual and mono-lingual interfaces. This was done through think-aloud protocols, observation, and interviews of users engaged in search tasks [3]. Their main finding is that their users did not make significant use of the cross-lingual functionalities of the system, nor did they think about language aspects when searching for an image. This again speaks to the necessity of careful design for a task which will better capture the complexity of a cross-lingual search task.

UNIPD also recruited users to be observed on-site, and constrained the task (in its first cycle) to require users to make a rapid decision of whether an image was relevant or not [2]. One of the conclusions pertinent to future cycles of the task is that the users are likely to be satisfied with a *similar* image, not necessarily needing the exact item designated correct by the game design. Designing future tasks might be well served in attempting to capture this usage-oriented aspect of user satisfaction.

The submission from the University of Westminster explored user's interaction with the facility provided by Flickling to add user-specific translation terms [6]. By exploring the user's perceived language skills and usage of the personal dictionary feature, experiments demonstrated that even with modest language skills, users were interacting with and using the dictionary-edit feature. Results point towards further study of collaborative translation in the global web space.

Finally, the group from IIIT-H studied the effects of language skills on user's search behaviour [7]. Results showed that user's typically started with a monolingual interface (the majority of

users having Spanish as their mother tongue) but soon moved to the cross-language interface, making use of facilities such as search hints when searching in languages other than their mother tongue. Overall, users mainly searched in their native language in which they felt more confident than searching (far less) in their passive languages. An interesting result was that, on average, users found more images successfully using the monolingual interface.

5 Conclusions

This paper has described a radical approach to studying user-orientated aspects of cross-language image search: iCLEF2008 has attempted 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). Although this initial attempt at encouraging greater participation in user-orientated evaluation resulted in submissions from 6 groups (the largest number of groups submitting to iCLEF in recent years), however a much larger number of users did make use of the system during the period of data collection showing potential for further experiments in 2009. The results of the experiments will be used to inform more usage-oriented tasks for future cycles; the methodology has proven to be lightweight and should be helpful for future participants; the logs will be a sustainable and reusable resource for future user-orientated studies of cross-language search behaviour.

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 this volume.

[2] Maria Di Nunzio, G.: "Interactive" Undergraduate Students: UNIPD at iCLEF 2008. In this volume.

[3] Vassilakaki, E., Johnson, F., Hartley, R.J., Randall, D.: A Study of Users' Image Seeking Behaviour in Flickling. In this volume.

[4] Karlgren, J. SICS at iCLEF 2008: User confidence and satisfaction inferred from iCLEF logs. In this volume.

[5] 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 this volume.

[6] Tanase, D. I. and Kapetanios, E.: Evaluating the impact of personal dictionaries for cross-language information retrieval of socially annotated images. In this volume.

[7] Vundavalli, S.: Mining the behaviour of users in a multilingual information access task. In this volume.

[8] Clough, P., Gonzalo, J., Kargren, J., Barker, E., Artiles, J. and Peinado, V.: Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge, In Proceedings of the Workshop on novel methodologies for evaluation in information retrieval, 30th European Conference on Information Retrieval, Glasgow, 30th March-3rd April 2008.