Creating Re-Useable Log Files for Interactive CLIR

Paul Clough
Department of Information Studies
University of Sheffield
Sheffield UK
p.d.clough@sheffield.ac.uk

Julio Gonzalo
UNED
C/ Juan del Rosal
16 28040 Madrid, Spain
julio@lsi.uned.es

Jussi Karlgren
SICS
Isafjordsgatan 22
120 64 Kista, Sweden
jussi@sics.se

ABSTRACT

This paper discusses the creation of re-useable log files for investigating interactive cross-language search behaviour. This was run as part of iCLEF 2008-09 where the goal was generating a record of user-system interactions based on interactive cross-language image searches. The level of entry to iCLEF was made purposely low with a default search interface and online game environment provided by the organisers. User-system interaction and input from users was recorded in log files for future investigation. This novel approach to running iCLEF resulted in logs containing more than 2 million lines of data.

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. 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 there are problems: user populations are 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 user interface.

In 2008 and 2009 iCLEF organisers decided to try something different: we provided a default multilingual search system (called `Flickling`) which accessed images from `Flickr`. Many images in `Flickr` contain metadata in multiple languages thereby providing a semi-realistic search scenario. For users of `Flickling` the task was kept very simple: given an image find it again. Users did not know in advance the languages in which the image was annotated; therefore searching in multiple languages was essential to get optimal results. The iCLEF interactive search task was publicised to attract as many users as possible from all around the world and the whole evaluation event was centered around an online: the more images found, the higher a user was ranked. This helped to encourage participation but also make the evaluation task engaging and addictive. The focus of iCLEF 2008-09 was on search log analysis rather than the more traditional focus of system design.

Interaction by users with the system was recorded in custom log files which were shared with iCLEF participants for further analyses. `Flicking` also gathered input from users such as their language skills, reasons for aborting a search task and comments on their search experience. These logs, a form of simulated interaction data[1], provide a resource for further study of interactive cross-language search behaviour. Our experiences in iCLEF have been positive and for the first time in this kind of evaluation setting we have been able to produce a re-useable output from interative experiments.

2. ICLEF METHODOLOGY

Our primary goal for iCLEF 2008-09 was to harvest a large search log of users performing multilingual searches on Flickr. Participants in iCLEF 2008-09 could perform two tasks: (1) analyse log files based on all participating users (default option) and, (2) execute their own interactive experiments with the interface provided by the organisers.

Generation of search logs. Participants could 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. Overall 435 subjects contributed to the logs.

Interactive experiments. Participants could 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 organisers provided assis-

[1]These logs can be downloaded from `http://nlp.uned.es/iCLEF/`

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

tance with defining appropriate user groups and image lists, for example, within the common search interface. Overall 6 groups participated in iCLEF and conducted experiments using `Flickling`.

2.1 Flickling

The `Flickling` system was created as a default search application for the evaluations [1]. The intention was to provide a standard baseline interface that was independent of any particular approach to cross-language search assistance. Functionality provided by the system included: user registration and recording of language skills, localisation of the interface, monolingual and multilingual search modes, automatic term-by-term query translation facility, query translation assistant allowing users to select/remove translations and add their own, query refinement assistant allowing users to refine or modify terms suggested by Flickr, control of game-like features and post-search questionnaires (launched after image found/failed, and a final questionnaire launched after the user had searched for 15 images). Customised logging recorded a rich amount of user-system interaction including explicit success/failure of searches, users' profiles, and post-search questionnaires for every search (over 7,500 questionnaires)

2.2 Search Task

The task was organised as an online game: the more images found, the higher a user was ranked. Depending on the image, the source and target languages, this could be a very challenging task. To have an adaptive level of difficulty, we implemented a hints mechanism. At any time whilst searching, the user was allowed to quit the search (skip to next image) or ask for a hint. The first hint was always the target language (and therefore the search became mono or bilingual as opposed to multilingual). The rest of the hints were keywords used to annotate the image. Each image found scores 25 points, but for every hint requested, there was 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. We therefore decided to remove time restrictions from the task definition.

2.3 Generated Logs

Overall, the logs collected and released during the iCLEF 2008 and 2009 campaigns contain more than 2 million lines. Table 1 summarises the most relevant statistics of both search logs. The log files record various user-system interactions such as qureries, results, items clicked, selected query translations, query modifications, feedback from users and navigational actions (e.g. next/previous page). In total 435 users contributed to the logs and generated 6,182 valid search sessions (a session is when a user logs in and carries out a number of searches). The logs provide a rich source of information for studying multilingual search from a user's perspective. iCLEF participants analysed the log files in various ways [2] [3]. For example, to discover actions leading to aborting a search task, investigating the effects of language skills on search behaviour, observing the switching behaviour of users between languages within a search session, investigating when users added translation terms to the `Flickling` dictionary and why, and the effects of ambigious search terms on search results and user behaviour.

Table 1: Statistics of iCLEF 2008/2009 search logs.

	2008	2009
subjects	305	130
log lines	1,483,806	617,947
target images	103	132
valid search sessions	5,101	2,410
successful sessions	4,033	2,149
unsuccessful sessions	1,068	261
hints asked	11,044	5,805
queries in monolingual mode	37,125	13,037
queries in multi-lingual mode	36,504	17,872
manually promoted translations	584	725
manually penalised translations	215	353
image descriptions inspected	418	100

3. DISCUSSION

What we have done in iCLEF is to observe the user-system interactions of users recruited to perform assigned tasks (real users/interactions, simulated tasks). The focus has moved from comparing different aspects of cross-language search assistance using more classical TREC (Interactive) style of experiment to using a default system and (simple) set of search tasks to provide a more realistic setting in which to conduct experiments and record and analyse user-system interactions. The data collection has not been constrained to subjects recruited by participating groups, but also involved recruiting subjects on an individual basis with the aim of contributing to the search log. This community (and game-like) approach to generating search logs is perhaps one way to generate resources that can be used by researchers and without the limitations and ethical concerns imposed when using logs from commercial web search engines.

However, although the user-system interaction logs provide a useful re-useable resource for studying user-system interaction and search behaviours, we also recognise the limitations of our approach. For example, the logs reflect only a specific known-item search task and participants experiment only with a single pre-defined search interface. In the future one could imagine using the log files to record behaviour for a specific version of the `Flicking` interface with systematic modifications carried out and user behaviours compared and used to evaluate various forms of search assistance.

4. REFERENCES

[1] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F. (2009) FlickLing: a multilingual search interface for Flickr, In Working Notes for the CLEF 2008 Workshop.

[2] Gonzalo, J., Clough, P. and Karlgren, J. (2009), Overview of iCLEF2008: Search Log Analysis for Multilingual Image Retrieval, In Proceedings of 9th Workshop of the Cross-Language Evaluation Forum (CLEF'08), September 17-19 2008, LNCS 5706, pp. 227-235.

[3] Gonzalo, J., Peinado, V., Clough, P. and Karlgren, J. (2009) Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy Environment, In Working Notes for the CLEF 2009 Workshop.