Trusting the results in crosslingual keyword-based image retrieval

Jussi Karlgren and **Fredrik Olsson**
SICS
Kista, Sweden
newtext@sics.se

Abstract

This paper gives a brief description of the starting points for the experiments the SICS team has performed in the 2006 interactive CLEF campaign.

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

trust, terminology, interactive information retrieval, cross-language information retrieval

Keywords

CLEF, iCLEF, Flickr, online photo sharing, multilingual image search, user studies, evaluation, evaluation campaigns

Keywords can be obscure

Flickr, the database chosen for this years interactive CLEF experiments, does not offer direct image search capabilities; searching for images in the database is done by searching keywords attached to the images by photographers or in some cases other viewers.

Searching for images using textual cues is a challenging task. There are many non-conventionalised keyword schemes, especially in a context such as Flickr, where keywords can be freely chosen. In Flickr, to compound the complexity of the task, the keywords frequently are given in several languages.

This means that we expect image viewers to be unsure of finding the right answer in Flickr.

Hypothesis 1: Viewers are uncertain of whether they have found everything that they set out to find.

Terms are related by their distribution

How could technology improve the sense of confidence viewers have in the results they have found? Our contention is that the descriptive qualities of a term are partially obscure to a user, and that this obscurity is exacerbated by cross-lingual effects - trying to use keywords in a foreign language will be more daunting than using terms in one's own. From other recent research we have investigated the effects of using distributional data to postulate semantic relations between terms [2] and in this experiment we make use of distributional similarity to provivde related terms. We display the terms used by the user, together with distributionally related terms in a word space screen for the user.

Hypothesis 2: By displaying the terminological space of tags related by distributional analysis to the ones used by the viewer we can encourage users to further search activity.

Relevance can be imprecise

Evaluation of user satisfaction is a contentious issue. The target notion of “Relevance” – related to and inspired by the everyday notion – in evaluating information retrieval systems is formalized to be an effective tool for focused research. Much of the success of information retrieval as a research field is owed to this formalization. But “Relevance” does not take user satisfaction or information need fulfilment into account. It is unclear how it could be generalized to the process of retrieving other than factual accounts. It is binary, where the intuitive and everyday understanding of relevance quite naturally is a gliding judgment. It does not does not take sequence of presentation into account - after seeing some information item, others may immediately lose relevance.

Trying to extend the scope of an information retrieval system so that it is more task-effective, more personalized, or more enjoyable will practically always carry an attendant cost in terms of lowered formal precision and recall as measured by relevance judgments. This cost is not necessarily one that will be noticed, and most likely does not even mirror a deterioration in real terms - it may quite often be an artefact of the measurement metric itself. Instead of being the intellectually satisfying measure which ties together the disparate and vague notions of user satisfaction, pertinence to task, and system performance, “Relevance” gets in the way of delivering all three.

The underlying hypothesis of our research activities is that the target notion for evaluating information access systems needs to cater for more than just topical relevance.

Hypothesis 3: User satisfaction is measurable and quantifiable in some respects.

Experiment

Twelve users have performed the three experimental tasks (as described in the track overview[1]):

- Topical ad-hoc retrieval: *Find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside.*
- Creative open-ended retrieval: *Find five illustrations to the article ”The story of saffron”*
- Visually oriented task: *What is the name of the beach where this crab is resting?* (along with a picture of a crab lying in the sand).

This year’s iCLEF did not contribute to the evaluation by requiring any specific metrics to be employed. Our evaluation has centered on the user experience of completion and satisfaction rather than accuracy. The three metrics we have employed are

1. Happy (all tasks): Are you satisfied with how you performed the task?
2. Complete (creative task, ad-hoc task): Did you find enough, or would you have continued if there had not been a time limit? How long would it have taken to make you satisfied, do you think?
3. Quality (creative task): Compare the illustrations with some given set. Are any of these better than your retrieved images?

As an additional metric, after a query has been performed, a terminology display of the terms used is displayed, where the terms employed by the user are displayed in order, together with other terms, related to the original terms by distributional metrics. The users can browse through the word space, and return to Flickr to rerun searches: we tabulate the number of additional searches made through this interface.

1 Conclusions

The conclusions and final analysis of our results will be presented at the workshop.

Acknowledgments

This work has been partially supported by the European Commission, through the DELOS Network of Excellence (FP6 G038-507618).

References

[1] Julio Gonzalo, Jussi Karlgren, and Paul Clough. iclef 2006 overview: Searching the flickr www photo-sharing repository. In *This volume*, 2006.

[2] Jussi Karlgren and Magnus Sahlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341, 2005.