CWI at the photo retrieval task of ImageCLEF 2009

Theodora Tsikrika and Arjen P. de Vries
CWI, Amsterdam, The Netherlands
Theodora.Tsikrika@cwi.nl, arjen@acm.org

Abstract

CWI's experiments investigate the usefulness of clickthrough data for improving the diversity of image retrieval results. We use the search logs provided to us by Belga to find relevant images; we consider that these correspond to images clicked for queries exactly matching or best matching a topic's title and cluster titles. To reduce the noise, we also filter these results and only consider those clicked images that are also retrieved by a text-based approach that uses the image captions. To promote diversity, we interleave the images retrieved in the previous step for each of the cluster titles (and also the title). However, given that the clickthrough data available to us cover only a small part of the collection used in the photo retrieval task, our experimental results are inconclusive, although they do provide indications on the reliability of using image search clickthrough data to identify relevant images.

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.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, diversity, clickthrough data, search logs, implicit feedback

1 Introduction

CWI participated in the photo retrieval task at ImageCLEF 2009 in the context of the research activities of the VITALAS[1] (Video and image Indexing and reTrievAl in the LArge Scale) project. One of the research directions investigated in this project is the usefulness of users' implicit feedback [6], in particular users' *clickthrough data* collected in search interaction logs, in a variety of applications, such as image annotation [12]. In VITALAS, the necessary resources for such experiments, i.e., the image collection and clickthrough data generated from users' interactions with this collection, have been made available by *Belga News Agency*[2], a picture portal that provides access to photographic images covering a broad domain. Given that this year's photo retrieval ImageCLEF task also uses images from the same content provider [10], we were motivated

[1] http://vitalas.ercim.org/

[2] http://www.belga.be/

to participate in this task and examine whether clickthrough data could be a useful source of evidence for improving the diversity of image retrieval results. Since this additional resource is not available to the participants of the photo retrieval task (only to VITALAS partners), the main aim of our participation has been to contribute retrieval results to the generation of the pools for the development of this test collection, and then to also examine how the effectiveness of our search log-based approaches compares to the more expensive clustering methods typically employed in such a task [1]. Section 2 presents the employed approaches that combine textual and search log-based evidence, Section 3 discusses the experimental results achieved by the officially submitted runs, and Section 4 concludes this paper.

2 Approach

Clickthrough data consist of the queries submitted by the users of a retrieval system, together with the documents in the retrieval results that these users selected to click on in response to their queries. Such implicit feedback is considered to "weakly" indicate the relevance of the document to the query for which it was clicked [3]. Using clickthrough data as a source of evidence of relevance is particularly attractive since they can be gathered in large quantities without any major effort on the part of content owners and without any explicit user intervention. Their major shortcoming, on the other hand, is that they are sparse, since they only cover the part of the collection that has been previously accessed, and potentially noisy. In the context of an image retrieval task that aims to promote diversity in the top n ranks, *image search clickthrough data* have the potential to minimise the tradeoff between precision and diversity, since they are both quite reliable as relevance assessments (even over 80% in some cases [11]), and they also include many query variations [9] that could be used to reflect the different topical facets of an information need.

The simplest approach to find **relevant** images using clickthrough data is to consider the images clicked for queries *exactly matching* the *title* (T) or the *cluster title(s)* (CT) of a topic (see the overview of the photo retrieval task [10] for details on the topic format and other aspects of the test collection). Given though that users with the same or similar information needs may submit different textual queries, an exact match approach may not be able to produce results for some of these topics; therefore, methods with less stringent matching criteria are also needed. To this end, we also employ approaches that consider as relevant those images clicked for queries that *best match* the title or the cluster title(s) of a topic. In particular, we find for each of these topic fields the images clicked for their best matching query which we consider to be the top ranking query using the NLLR (normalized log-likelihood ratio) retrieval model [7], a simple derivation of a language model with linear interpolation smoothing [4] that produces log-linear scores. In summary, we construct, for each field of each topic (i.e., the topic's title T and each of its cluster titles CT_i, $i = 0, 1, ..., m$), a list of images clicked for its exactly matching query and a list of images clicked for its best matching query, and we rank the images within each individual list by their click count; that means that for each topic we construct $2 * (1 + m)$ ranked image lists.

To eliminate some of the false positives in these search log-based rankings, i.e., images that have been clicked without being relevant to the submitted query, we also take into account the textual evidence in the form of captions that accompany the images in the collection. We index the images in the collection using their captions, after applying stemming and stopword removal. We use the title field of each topic as the query and NLLR as the retrieval model to produce a top 1000 ranking. Then, by keeping in the search log-based rankings generated in the previous steps only the images that also appear in the top 1000 text-based ranking, a potentially less noisy result is produced for each of the considered topic fields.

To promote **diversity** among the top ranked retrieval results, we simply interleave the images in the lists produced for each topic in the previous steps, i.e., either the lists containing the images clicked for the query exactly matching each of the cluster titles and/or the title of the topic, or the lists containing the images clicked for the query best matching each of the cluster titles and/or the title of the topic, or the lists containing the images that both were clicked for the query exactly matching each of the cluster titles and/or the title of the topic and also appear in the top 1000 text-based retrieval results.

The approaches described above are listed below in a more succinct form (the ones marked by an asterirk (*) correspond to runs officially submitted to the task):

1. **cwi1_T_TXT**: Images clicked for the query exactly matching the title.
2. **cwi2_CT_TXT**: Interleave the m lists of images clicked for the queries exactly matching each of the cluster titles.
3. **cwi3_TCT_TXT***: Interleave the $m + 1$ lists of images clicked for the queries exactly matching the title and each of the cluster titles.
4. **cwi4_T_TXT**: Images clicked for the query best matching the title.
5. **cwi5_CT_TXT**: Interleave the m lists of images clicked for the queries best matching each of the cluster titles.
6. **cwi6_TCT_TXT***: Interleave the $m+1$ lists of images clicked for the queries best matching the title and each of the cluster titles.
7. **cwi7_T_TXT**: Images retrieved in the top 1000 of a text-based approach and also clicked for the query exactly matching the title.
8. **cwi8_CT_TXT***: Interleave the m lists that contain images retrieved in the top 1000 of a text-based approach and also clicked for the queries exactly matching each of the cluster titles.
9. **cwi9_TCT_TXT***: Interleave the $m+1$ lists that contain images retrieved in the top 1000 of a text-based approach and also clicked for the queries exactly matching the title and each of the cluster titles.

The approaches presented thus far find the relevant images in the search logs by considering those previously clicked for a query that matches a topic field, while, for the promotion of diversity, they require that the different topical facets of an information need, corresponding to the different clusters of a topic in the photo retrieval task, are known in advance. We also investigate an approach that does not require the availability of such information[3]. This approach, denoted as **cwi10_T_TXT***, first finds the queries for which the images retrieved in the top 1000 by the text-based approach have been previously clicked. Then the images clicked for each these queries are found in the clickthrough data, ranked by their click count, and approach *cwi2_CT_TXT* is applied with each of these queries treated as a cluster title.

The clickthrough data used in our experiments have been extracted from Belga search logs that correspond to a 15-month period (June-September 2007 and January-December 2008). However, these data cover a time period that it is slightly different to the time period within which the Belga images that comprise the ImageCLEF photo collection have been made available online to the Belga users. Therefore, these logs do not record many of the interactions with these particular images and thus are quite sparse with respect to this dataset. Figure 2 presents, for each topic, the images that have been previously clicked for the topic's title and cluster titles (both the total number of clicked images and also those that also belong to the top 1000 retrieved using the text-based approach). For 11 out of the 50 topics, there are no clicked images for queries exactly matching either the title or the cluster title(s), whereas for 8 more topics, there is no overlap between the clicked images found and those retrieved in the top 1000 by the text-based approach.

Overall, the number of images found in the search logs for this data set is very low (even less than 10 in many cases), which makes it very difficult to evaluate the usefulness of the proposed approaches. Nevertheless, we decided to submit our runs so as to get an indication of their effectiveness. However, to ensure that we submit a sufficient number of results for each of the

[3] This is actually one of the objectives examined in the context of this task, since only half of the 50 available topics contain cluster information, whereas the other half only contain a title [10].

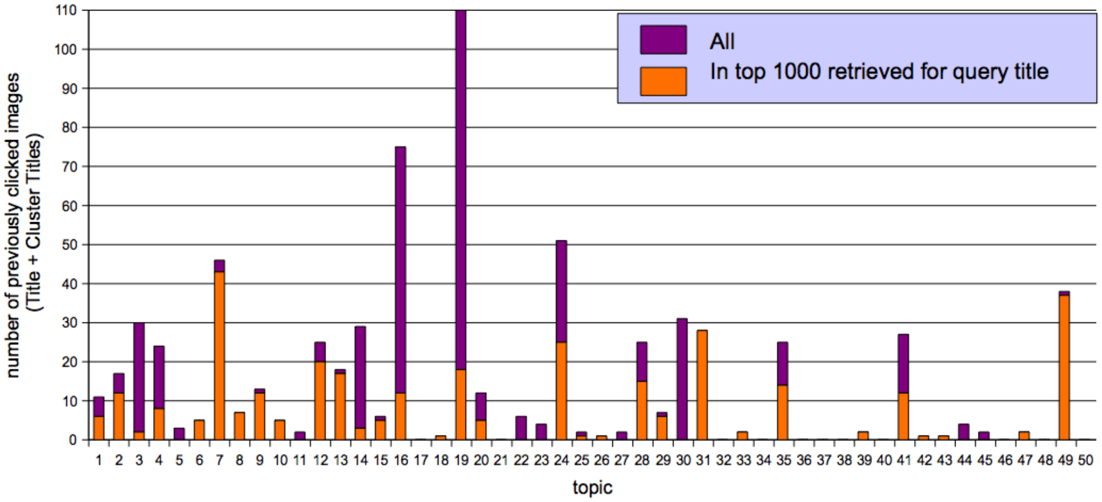

Figure 1: The number of images that have been previously clicked for each topic's title and cluster titles (both the total number of clicked images and also those that also belong to the top 1000 retrieved using the text-based approach).

topics, we append to each ranking the top 100 images retrieved by the text-based approach (taking care to remove from these top 100 those already retrieved by the search log-based approaches).

We conducted our experiments using PF/Tijah[4] [5], a research project run by the University of Twente, which aims at creating a flexible environment for setting up search systems. PF/Tijah is part of the open source release of MonetDB/XQuery[5], which is being developed in cooperation with CWI, Amsterdam and the University of München. PF/Tijah combines database and Information Retrieval technologies by integrating the PathFinder (PF) XQuery[6] compiler [2] with the Tijah XML retrieval system [8].

3 Results

Table 1 presents the results of our official submissions. Out of the 5 submitted runs, the effectiveness of the 4 best performing ones is comparable, whereas run *cwi10_T_TXT* is not very effective as it is probably affected by topic drift. The mean P@10 and P@20 of the 4 best performing runs is around 0.75 over the 50 topics and also over each of the two subsets. This indicates that both the search log-based and the text-based approach, and in particular their combination (i.e., runs *cwi9_TCT_TXT* and *cwi8_CT_TXT*) are effective in identifying relevant images. A further examination of the P@10 values achieved by the best performing *cwi9_TCT_TXT* for each of the topics (as these P@10 values are plotted against the number of top ranked images that have also been previously clicked in Figure 3) indicates that the majority of topics (35/50) perform better than the mean P@10 (the median P@10 is 0.8). Figure 3 further indicates the high reliability of clicks as relevance assessments.

Regarding the diversity of the retrieval results, our 4 best performing runs manage to retrieve, in the top 10 ranks, images belonging to at least half of the different topical facets that have been considered in the evaluation. The results regarding the CR@10 are slightly better in the second part of the topics (topics 26-50). Given that these topics do not contain cluster titles and thus all the approaches that rely only on them do not produce any results and simply consider the top

[4] http://dbappl.cs.utwente.nl/pftijah/

[5] http://www.sourceforge.net/projects/monetdb/

[6] http://www.w3.org/TR/xquery/

100 images retrieved by the text-based approach that have been appended, these results indicate that this text-only approach and also the images clicked for the title achieve quite a high diversity. However, a per-topic analysis of the CR@10 (and F-measure) values is required so as to reach more reliable conclusions.

Table 1: Results for the CWI official submissions to the photo retrieval task at ImageCLEF 2009.

overall	relative	run	F-measure	CR@10	CR@20	P@10	P@20	$rel \wedge retr$
rank	rank							
All Topics (relative ranking among modality TXT)								
33	11	CWI9_TCT_TXT	0.6467	0.5607	0.6355	0.76	0.78	3914
42	18	CWI3_TCT_TXT	0.6301	0.5639	0.6292	0.71	0.74	4179
45	20	CWI8_CT_TXT	0.6217	0.5250	0.6305	0.76	0.78	3834
53	24	CWI6_TCT_TXT	0.6033	0.5027	0.6140	0.75	0.78	3738
73	38	CWI10_T_TXT	0.4445	0.4854	0.5787	0.41	0.40	7105
Topics 1-25 (relative ranking among all runs)								
33	46	CWI9_TCT_TXT	0.6255	0.5296	0.6292	0.76	0.78	1992
42	48	CWI3_TCT_TXT	0.6216	0.5515	0.6242	0.71	0.74	2229
45	53	CWI8_CT_TXT	0.6017	0.4962	0.6159	0.76	0.77	1965
53	61	CWI6_TCT_TXT	0.5633	0.4518	0.5828	0.75	0.77	1868
73	73	CWI10_T_TXT	0.4455	0.4794	0.5813	0.42	0.40	3879
Topics 26-50 (relative ranking among all runs)								
33	21	CWI9_TCT_TXT	0.6670	0.5918	0.6418	0.76	0.78	1922
53	40	CWI6_TCT_TXT	0.6407	0.5537	0.6451	0.76	0.78	1870
45	41	CWI8_CT_TXT	0.6407	0.5537	0.6451	0.76	0.78	1869
42	42	CWI3_TCT_TXT	0.6385	0.5762	0.6342	0.72	0.73	1950
73	69	CWI10_T_TXT	0.4434	0.4914	0.5761	0.40	0.41	3226

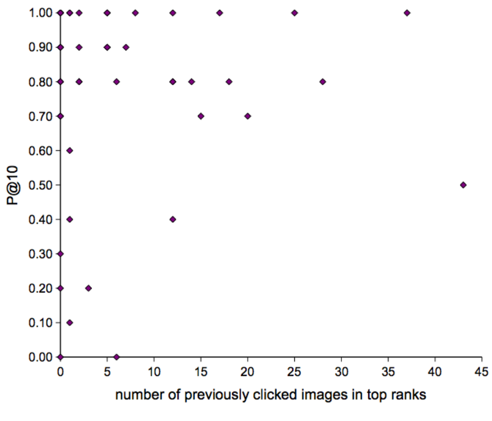

Figure 2: P@10 values for run *cwi9_TCT_TXT* plotted against the number of images that have been previously clicked and are also retrieved in the top 1000 of the text-based NLLR approach.

4 Conclusions

We investigated the effectiveness of employing clickthrough data to improve the diversity among the top ranked retrieval results. Unfortunately the Belga search logs available to us at the time of submission cover only a really small part of the image collection used in the photo retrieval task. This renders our experimental results inconclusive, although they do indicate the reliability of using image search clickthrough data as relevance assessments. We are currently performing further experiments using an additional sample of Belga search logs and are also exploring approaches to deal more effectively with the sparseness of clickthrough data. In the future, we also aim to apply more principled approaches in the diversification of image retrieval results.

5 Acknowledgements

This material is based on work supported by the European Union via the European Commission project VITALAS (contract no. 045389).

References

[1] T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Advances in Multilingual and Multimodal Information Retrieval: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*, Lecture Notes in Computer Science. Springer, 2009.

[2] P. Boncz, T. Grust, M. van Keulen, S. Manegold, J. Rittinger, and J. Teubner. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 479–490, 2006.

[3] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 239–246, 2007.

[4] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL 1998)*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584. Springer, 1998.

[5] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. PF/Tijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR) (held in conjunction with SIGIR 2006)*, pages 12–17, August 2006.

[6] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[7] W. Kraaij. *Variations on Language Modeling for Information Retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2004.

[8] J. List, V.Mihajlović, G.Ramírez, A. de Vries, D. Hiemstra, and H. Blok. TIJAH: Embracing IR Methods in XML Databases. *Information Retrieval*, 8(4):547–570, 2005.

[9] M. L. Paramita, M. Sanderson, and P. Clough. Development of a collection to support diversity analysis. In *Proceedings of the Workshop on Redundancy, Diversity, and Interdependent Document Relevance (IDR 2009) - held at the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, 2009.

[10] M. L. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the ImageCLEFphoto task 2009. In *CLEF working notes 2009*, 2009.

[11] G. Smith and H. Ashman. Evaluating implicit judgements from image search interactions. In *Proceedings of the Web Science Conference: Society On-Line (WebSci 2009)*, 2009.

[12] T. Tsikrika, C. Diou, A. P. de Vries, and A. Delopoulos. Image annotation using clickthrough data. In *Proceedings of the 8th International Conference on Content-based Image and Video Retrieval (CIVR 2009)*, 2009.