Experiments with Clustering the Collection at ImageCLEF 2007

Osama El Demerdash, Leila Kosseim and Sabine Bergler
Concordia University
osama_el,kosseim,bergler@cse.concordia.ca

Abstract

We present our participation in the 2007 ImageCLEF Ad-hoc photographic retrieval task. Our first participation in this year's imageCLEF comprised six runs. The main purpose of three of these runs was to evalute the text and visual retrieval tools as well as their combination in the context of the given task. The other purpose of our participation is to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of ImageCLEF AD-hoc task. We use the preclustered collection to augment the search results of the retrieval engines. For retrieval we used two publicly available libraries; *Apache Lucene* for text and *LIRE* for visual retrieval. The clustered-augmented results reduced slightly the precision of the initial runs. While the aspired results have not yet been achieved, we note that the task is useful in assessing the validity of the clusters.

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 Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image Retrieval, Clustering

1 Introduction

We present our participation in the 2007 ImageCLEF Ad-hoc photographic retrieval task. The task deals with answering 60 queries of variable complexity from a repository of 20,000 photographic images in the IAPR collection. A full description of the task and the collection can be found in [7]. Our first participation in this year's imageCLEF comprised six runs. The main purpose of three of these runs was to evaluate the text and content-based retrieval tools in the context of the given task. We therefore would like to stress that the evaluation of these tools can only be considered under the given parameters of the task, including the queries, the image collection and our utilization of these tools.

The other purpose of our participation is to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of ImageCLEF AD-hoc task. We also want to advance the hypothesis that while this task of ImageCLEF might not be best

suited for the evaluation of interactive methods (despite the designation of runs as Manual/Auto), it could still be useful in the evaluation of certain aspects of such methods such as the validity of the initial clusters in our case.

2 Related Work

Clustering, as an unsupervised machine learning mechanism, has rarely been investigated within the context of the ImageCLEF Ad-hoc Retrieval task. This could be due to that clustering methods lend themselves more readily to interactive tasks and iterative retrieval. In the IR field, Clustering has been experimented with extensively [10]. Its different applications involve clustering the whole data collection, part of it or the search results. In [11], images are clustered using labels from the surrounding HTML text. [3] applied clustering to content-based image retrieval using the Normalized Cut (NCut) algorithm under a graph representation. Another spectral clustering algorithm *Locality Preserving Clustering (LPC)* was introduced in [12] and found to be more efficient than NCut for image data.

3 Resources

For retrieval we used two publicly available libraries; *Apache Lucene* for text and *LIRE* for visual retrieval. Since our runs involved only English/English and Visual queries we did not make use of any translation tools.

3.1 Text Retrieval

For text retrieval we used the Apache Lucene engine [6] which implements a TF-IDF paradigm. Stop-words were removed and the data was indexed as *field data* retaining only the *title*, *notes* and *location* fields all of which were concatenated into one field. This helped reduce the size of the index, since our initial plan was to base the clustering on word-document cooccurence and document-document similarity matrices. The number of indexed terms is 7577 from the 20,000 English source documents. All text query terms were joined using the *OR* operator. We did not apply any further processing of text queries.

3.2 Content-based Retrieval

For visual retrieval, we employed v0.4 of the LIRE library which is part of the Emir/Caliph project [9] available under the Gnu GPL license. At the time of carrying out the experiments LIRE offered three indexing options from the MPEG-7 descriptors: ScalableColor, ColorLayout and EdgeHistogram (a fourth one, Auto Color Correlogram, has since been implemented). We used all three indices. The details of these descriptors can be found in [4]. Only the best 20 images of each visual query were used. The visual queries consisted of the three images provided as example results. Thus, a maximum of 60 image results from visual queries were used in the evaluation.

4 Clustering Methodology

Three of our runs utilized preclustering of the data collection to augment the result set of the retrieval engines. Although we intended in the beginning to cluster the results obtained from the text retrieval and content-based retrieval, we resorted to clustering the collection, given the small number of relevant results per query (compared to results from searching the WWW for example).

We employed a simple one-pass clustering algorithm which relied on forming clusters of the terms in the documents as they are processed. If a document's similarity to a cluster exceeded a certain threshold (N), this document and its new terms are added to the term/document cluster.

If a document is not associated with any cluster, it is temporarily assigned its own, which is deleted in the end if no other documents are associated with it. Also clusters larger than size (S) were discarded. We did not however experiment with the parameters S and N and chose them with the little intuition we had about the data. The resulting clusters overlap and do not cover all documents.

For augmenting the results from the clusters, we searched the clusters for each result and whenever one was found we inserted the other members of the cluster at this position in the result set, taking care not to include duplicate results from different clusters.

5 Results and Analysis

Table 1 shows the results our runs obtained at ImageCLEF 2007. Our highest ranked run (clacTXCB) is the one that combined results from Lucene (text retrieval) and LIRE (visual retrieval), getting a higher MAP as well as better performance on all other measures than the other runs. For this run we used a combined list of the results from both engines ranking common results highest on the list. The relatively lower GMAP that all our runs obtained reflects the fact that we did not employ any semantic processing of the queries. Our simple method of augmenting results from preclustered data resulted in slightly worse results in all three cases: text, visual and their combination. The main reason is that our clusters were less fine-grained than the requirements of the queries.

Run ID	Modality	MAP	P10	P20	P30	BPREF	GMAP	REL_RETR
clacTXCB	MIXED	0.1667	0.2750	0.2333	0.2067	0.1599	0.0461	1763
clacCLSTXCB	MIXED	0.1520	0.2550	0.2158	0.1939	0.1445	0.0397	1763
clacTX	TEXT	0.1355	0.2017	0.1642	0.1617	0.1231	0.0109	1555
clacCLSTX	TEXT	0.1334	0.1900	0.1575	0.1522	0.1205	0.0102	1556
clacCB	Visual	0.0298	0.1000	0.1000	0.1033	0.0584	0.0058	368
clacCLSTCB	MIXED	0.0232	0.0817	0.0758	0.0722	0.0445	0.0038	386

Table 1: Results at ImageCLEF 2007

6 Conclusion and Future Work

We intend to experiment with clustering the result set as well as introducing query expansion and pseudo-relevance feedback. Our final target is clustering based on both text and visual features. There is very little evidence in the literature on clustering using both content-based and text-based features. [2] and [1] describe successive clustering applied on text features then image features. The textual features comprised a vision based text segment as well as the link information while the *Color Texture Moments (CTM)*, a combined representation of color and texture were chosen for visual features.

The only research we came across in the literature combining simultaneously image and textual features were from Microsoft Research Asia in 2005. [8] and [5] both use co-clustering techniques

Our first participation at ImageCLEF was satisfactory in that we were able to evaluate the IR tools we chose ,as well as the validity of the initial clusters produced from a simple unsupervised clustering method. We understand however that this task is not most suited for evaluating iterative retrieval, especially when it comes to usability factors. We look forward to participating in future years.

References

[1] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959, New York, NY, USA, 2004. ACM Press.

[2] Deng Cai, Xiaofei He, Wei-Ying Ma, Ji-Rong Wen, and HongJiang Zhang. Organizing www images based on the analysis of page layout and web link structure. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan*, pages 113–116. IEEE, 2004.

[3] Yixin Chen, James Z. Wang, and Robert Krovetz. Content-based image retrieval by clustering. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, New York, NY, USA, 2003. ACM Press.

[4] José Martínez (ed.). Mpeg-7 overview (version 10). Technical Report N6828, ISO/IEC JTC1/SC29/WG11 (MPEG), October 2004. online August 17, 2007 http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[5] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 112–121, New York, NY, USA, 2005. ACM Press.

[6] Otis Gospodnetic' and Erik Hatcher. *Lucene in Action*. 2005.

[7] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[8] Zhiwei Li, Gu Xu, Mingjing Li, Wei-Ying Ma, and Hong-Jiang Zhang. Grouping www image search results by novel inhomogeneous clustering method. In Yi-Ping Phoebe Chen, editor, *11th International Conference on Multi Media Modeling (MMM 2005)*, pages 255–261. IEEE Computer Society, 2005.

[9] Mathias Lux and Michael Granitzer. Retrieval of mpeg-7 based semantic descriptions. In *BTW-Workshop WebDB Meets IR at the GI-Fachtagung fr Datenbanksysteme in Business, Technologie und Web, University Karlsruhe, March 2005*, 2005.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. Online 17/08/2007. http://www-csli.stanford.edu/ schuetze/information-retrieval-book.html.

[11] Wataru Sunayama, Akiko Nagata, and Masahiko Yachida. Image clustering system on www using web texts. In *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, pages 230–235, 2004.

[12] Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 885–891, New York, NY, USA, 2004. ACM Press.