Using Visual Linkages for Multilingual Image Retrieval

Masashi Inoue

National Institute of Informatics, Tokyo, Japan
m-inoue@nii.ac.jp

Abstract. The use of visual features in text-based ad-hoc image retrieval is challenging. Visual and textual information have so far been treated equally in terms of their properties and combined with weighting mechanisms for balancing their contributions to the ranking. The use of visual and textual information in a single retrieval system sometimes limits its applicability due to the lack of modularity. In this paper, we propose an image retrieval method that separates the usage of visual information from that of textual information. Visual clustering establishes linkages between images and the relationships are later used for the reranking. By applying clustering on visual features prior to the ranking, the main retrieval process becomes purely text-based. Experimental results on the ImageCLEFphoto ad-hoc task show this scheme is suitable for querying multilingual collections on some search topics.

1 Introduction

In the case of annotation-based multilingual image retrieval without any translation, the target collection is limited to the images annotated in the query language. Even if users are aware of that there are many relevant images annotated in different languages, many may not spare any of their time to explore them and instead may consider a portion of images annotated in one language as sufficient. Through automation, cross-language information retrieval (CLIR) techniques may be of help by expanding the range of images accessible on some search topics.

Text-based or annotation-based retrieval is thought to be the basis of image retrieval because associated text information plays an important role in assuring the effectiveness of retrieval. However, annotation-based approaches have limitations due to the lack of sufficient textual annotations. One reason for this insufficiency is the high cost of creating annotations. Also, for most nonspecialists, it is not easy to imagine the need for annotating images for future use. Therefore, the question lies in how we can supplement textual information in existing image collections. To alleviate the problem of text scarcity in image retrieval, we propose to use a "knowledge injection" framework. The definition of "knowledge" in this paper is that the entities used correspond to human conceptualization and their relationships are intuitively understandable, often in the form of qualitative values. In addition, the segregation of provided and acquired

Table 1. Comparison of knowledge injection and information integration frameworks

Type of entity	Automatically generated	Manually created
Interpretable	Extracted knowledge injection	Provided knowledge injection
Non-interpretable	Data integration	Not defined

knowledge is important. From the main retrieval system's side, knowledge bases that are the result of manual effort are called "provided". In contrast, information automatically extracted from the target collection or other information source is called "acquired". Table 1 illustrates the relationship between knowledge injection and data integration, as well as the difference between provided and acquired knowledge. An example of provided knowledge is the WordNet thesaurus that injects the knowledge of relationships between words [1]. It has been applied to annotation-based image retrieval [2]. We investigate the use of acquired knowledge based on visual features of images in a target collection.

The use of visual features is conducted in the framework of a "find similar" task. This procedure can be executed in two ways. The first way involves searching for similar images after the initial result has been retrieved and sometimes makes use of user feedback. An on-line method for image retrieval involves clustering the retrieved results. For example, Chen et al. used a clustering method, which is originally developed for content-based image retrieval, in the annotation-based image retrieval framework, as post-processing after querying [3]. The second way involves finding similar image pairs or groups prior to querying. We investigate this second option in this paper. This choice is based on considerations of efficiency. The similarity calculation between images based on visual features is usually heavier than the one based on textual features. Thus, for image retrieval, it is sometimes desirable to conduct such computations off-line.

In the following sections, we introduce the idea of micro-clustering preprocessing to extract visual knowledge and describe the configuration of our retrieval models. We then show and analyze retrieval results for the ImageCLEFphoto collection. Finally, we conclude the paper.

2 System Description

2.1 Micro-Clustering

Our retrieval strategy consists of three distinct steps. The first step is the knowledge extraction by using micro-clustering. Two types of clustering can be imagined. One is macro-clustering, or global partitioning, when the entire feature space is divided into sub-regions. For the document collection $\mathcal{D}$, in the macro-clustering, the set of clusters $\mathcal{C}_g = \{c_1, c_2, \ldots, c_M\}$ are constructed where $\mathcal{D} = \bigcup_m c_m$ holds. Also, all documents belong to only one cluster. That is, if

$d_i \in c_m$, $\forall j \neq i$, $d_j \notin c_m$ and $\forall i$, $d_i \in c_m (m = 1, \ldots, M)$. The other is micro-clustering, or local pairing, where nearby data points are linked so that they form a small group in a particular small region of the feature space. The set of clusters $\mathcal{C}_l = \{c_1, c_2, \ldots, c_M\}$ are constructed and there is not any constraints as above. That is, each document does not have to be a cluster member and can be a member of more than one cluster. Usually the number of clusters M in micro-clustering is bigger than macro-clustering and the size of individual clusters are far smaller. In our system, micro-clustering was used to group images based on their visual similarities.

The concept of the micro-clustering has been developed to enhance the indexing of textual documents [4]. In [4], micro-clusters are treated as documents and used for classification. We used the same concept but with the features and similarity measure for visual information. Further, in this paper, the micro-clusters are regarded as the linkage information and used for re-ranking.

The process of clustering is as follows. First, visual features are extracted from all images. Simple color histograms are used. Since the images are provided in true color JPEG format, the histograms are created for the red (R), green (G), and blue (B) elements of the images. This results in three vectors for each image: $\mathbf{x}_r$, $\mathbf{x}_g$, and $\mathbf{x}_b$. The length of each vector, or the size of the histogram, $i = 256$. These vectors are concatenated and define a single feature matrix for each image: $X = [\mathbf{x}_r, \mathbf{x}_g, \mathbf{x}_b]$. Thus, the size of the feature matrix is i by j where $j = 3$.

The similarities between images are calculated using the above feature values. The similarity measure employed was the two-dimensional correlation coefficient r between the matrices. Assuming two matrices A and B, the correlation coefficient is given as

$$r = \frac{\sum_i \sum_j (A_{ij} - \bar{A})(B_{ij} - \bar{B})}{\sqrt{(\sum_i \sum_j (A_{ij} - \bar{A})^2)(\sum_i \sum_j (B_{ij} - \bar{B})^2)}}$$

where $\bar{A}$ and $\bar{B}$ are the mean values of A and B respectively.

Next, a threshold is set that determines which two or more images should belong to the same cluster. In other words, image pairs whose r score is larger than the threshold are considered identical during retrieval. At this stage, the threshold value is manually determined by inspecting the distribution of similarity scores so that relatively small numbers of images constitute clusters. Small clusters containing nearly identical images are preferred since visual similarity does not correspond to semantic similarity; however, visual identity often corresponds to the semantic identity.

2.2 Initial Retrieval

The novelty of our method in the ImageCLEF2006 is solely in the pre-processing of the retrieval. For the second step of the process, we used an existing search

engine, the Lemur Toolkit[1]. We used a unigram language-modeling algorithm for building the document models, and the Kullback-Leibler divergence for the ranking. The document models were smoothed using Dirichlet prior [5].

2.3 Re-ranking

The third step is the injection of the linkage knowledge extracted in the first step. The ranked lists given by the retrieval engine are re-ranked by using the cluster information. The ranked list is searched from the top and when an image that belongs to a cluster is found, all other members of the cluster are given the same score as the highly ranked one. This process is continued until the number of images in the list exceeds the pre-specified number, which is 1000 in our study.

3 Experiments

3.1 Experimental Configuration

The details of the test collection used are given in [6]. There are 20,000 images annotated in English and in German. Instead of viewing the collection as a single bilingual collection, it is regarded as a collection of 20,000 English images and a collection of 20,000 German images. Each annotation has seven fields but only the title and description fields were used.

For the comparison, six runs were tested for the monolingual evaluation. The query languages were English, German, and Japanese. The collection languages were English and German. We applied query translation. The Systran machine translation (MT) system[2] was used. Because of lack of direct translation functionality between German and Japanese in the MT system, English was used as the pivot language when querying German collections using Japanese topics. That is, Japanese queries were first translated into English, and then the English queries were translated into German. The relationship between query and document languages and monolingual runs' names is summarized in Table 2.

In the table, names of runs are assigned according to the following rules. The first element *mcp* comes from the proposed method, micro-clustering pre-processing, and represents our group. The second element *bl* indicates that the baseline method was used. When the micro-clustering pre-processing was used, the value of the threshold is used for this element. For example, *09* denotes that the pairs with correlation coefficients greater than 0.9 form a cluster. The next element concerns the query language and the fields of the search topics. Runs using English queries with only title fields are marked as *eng_t*. Similarly, the next element is the collection language and the fields of the annotations. Runs using the German collection with title and description fields are marked by *ger_td*. When half of the English collection and half of the German collection are mixed together, the notation is *half_td*, as shown in Table 3. The last element is

[1] http://www.lemurproject.org/
[2] http://babelfish.altavista.com/

Table 2. Summary of runs on monolingual collections (run names and MAP scores)

Query	Document Language			
Language	English		German	
English	mcp.bl.eng_t.eng_td.skl_dir	0.1193	mcp.bl.eng_t.ger_td.skl_dir	0.0634
German	mcp.bl.ger_t.eng_td.skl_dir	0.1069	mcp.bl.ger_t.ger_td.skl_dir	0.0892
Japanese	mcp.bl.jpn_t.eng_td.skl_dir	0.0919	mcp.bl.jpn_t.ger_td.skl_dir	0.0316

Table 3. Summary of runs on the linguistically heterogeneous collection (run names and MAP scores)

Query	Half English Half German Document Collection			
Language	Without pre-processing		With pre-processing	
English	mcp.bl.eng_t.half_td.skl_dir	0.0838	mcp.09.eng_t.half_td.skl_dir	0.0586
German	mcp.bl.ger_t.half_td.skl_dir	0.0509	mcp.09.ger_t.half_td.skl_dir	0.0374

the configuration of the retrieval engine. The simple Kullback-Leibler divergence measure was used for ranking (*skl*) and Dirichlet prior used for smoothing (*dir*). All runs used the same configuration.

3.2 Results for Homogeneous Collections

In the baseline runs, the collection language is the determining factor of retrieval performance as shown in Table 2. Searching an English collection is better in any query language. Furthermore, the translated topics from German to English on the English collection worked better than mono-lingual German topics on the German collection. The results for Japanese topics on the German collection were poor, because of poor machine translation.

3.3 Results for Heterogeneous Collections

The advantage of the visual-similarity based pre-clustering becomes clear from the application of linguistically heterogeneous image collections. Therefore, a linguistically heterogeneous collection was constructed by taking $10,000$ randomly chosen images from the English collection and the remaining $10,000$ images from the German collection. There were no overlapping images. Both English and German queries without translation were tested on this single collection with micro-clustering. Table 3 shows the result. Because half of relevant images are now in annotated in another language, the MAP scores were worse than mono-lingual cases. It was observed that the pre-processing gave no improvement in terms of the mean average precision (MAP) scores.

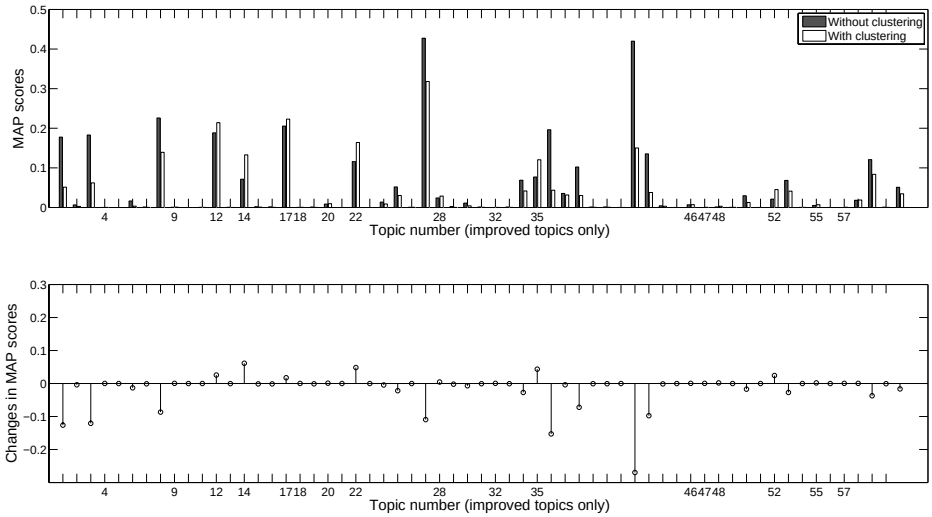

Fig. 1. MAP scores of each topic with and without pre-processing

When we looked at the results topic-by-topic, we found improvements for some topics, as shown in Figure 1. This suggests the need for topic analysis when applying micro-clustering pre-processing.

3.4 Analysis of Clustering Result

The generated clusters were small and often of size two: a cluster formed by a pair of images. We intended this to be the result of micro-clustering; we wanted quite small yet highly condensed clusters. The statistics of cluster sizes are as follows: mean = 12.72, standard deviation = 43.81, minimum = 0, and maximum = 368. Some clusters have more than 100 members. Such non-microclusters are not ideal because when one of their members appears in the list, the cluster dominates the entire list after re-ranking. Thus, clusters bigger than 6 were truncated to size 6.

3.5 Discussion

Incorporating visual pre-processing did not improve the average performance for all topics. This failure might be because clusters of irrelevant images were used rather than relevant ones. Because not all of the initially retrieved images were relevant, we may need to use certain tactics to select only highly relevant images. Also, there is a trade-off between the quality of clustering and the degree of search target expansion. In the experiment, the threshold value might be conservative

in order to avoid the inclusion of noisy clusters. More investigations are needed to clarify the effect of the threshold values.

The potential advantages of our approach over the usual query translation methods are as follows. First, there is no need to combine rankings given by multiple translated queries. Because the rank aggregation is difficult in IR, trial and error in the design of the merging strategies can not be avoided. Our approach outputs one ranking and the merging is not needed. Second, the systems do not have to deal with the languages. The method can be used even when the language distribution within the collection is unknown.

The limitations of our experimental setting should be noted. The test collection is built upon a random selection from two language collections. Thus, nearly identical images that might have been originally created in a sequential manner could have been split into two languages. However, in reality, many similar image pairs may have annotations in only one language. For example, if one photographer took photos of an object, it would be natural to assume that all of these photos would be annotated in the same language. In the future, we will investigate more realistic linguistically heterogeneous collections.

In regard to the generalizability of the outcome of our experiment, the properties of the target collections is a big issue. The IAPR TC-12 collection used here mostly contains tourist photos. It contains a moderate amount of nearly identical images. If the target collection contained a larger amount of nearly identical images, the gain given by the pre-processing may be higher than it was with IAPR TC-12, whereas if it had few similar images, the preprocessing would not likely improve the retrieval effectiveness. Another issue is that the some images of a topic mostly exist in the sphere of one language. We do not have enough knowledge about the relationship between language and search topics yet.

4 Conclusion

This paper presented the results of experimental runs on the ImageCLEF2006 ad-hoc photo retrieval task. The goal was to investigate the possibility of a modular retrieval architecture that uses visual and textual information at different stages of retrieval. A visual feature-based micro-clustering was used for the linkage of nearly identical images annotated in different languages. After this preprocessing, the retrieval was conducted as a monolingual retrieval using query language. Then, images that are linked to the highly ranked images are pulled up. As a result, images annotated in different languages can be searched beyond the language barriers. In the experiment, although the mean performance over all topics did not improve, the individual average precision for some topics improved. The gains originated from the inclusion of additional images annotated in another language to the ranked list.

The biggest issue remaining is the lack of understanding of real-world needs for the cross-language image access. It is not fully known in what sort of search task that cross-language retrieval techniques will be helpful in information access in general [7]. Considering the language-independent nature of visual represen-

tation, cross-language image retrieval may be one such task. However, even in image retrieval, it is not clear how we can characterize the target collection from the perspective of image type and linguistic non-uniformity. Besides the refinement of pre-processing and post-processing, our future work will include developing a methodology for analyzing tasks and collections.

Acknowledgment

This research is partly supported by MEXT Grant-in-Aid for Scientific Research on Priority Areas (Cyber Infrastructure for the Information-explosion Era) and Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (17700166).

References

1. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38** (1995) 39–41
2. Smeaton, A.F., Quigley, I.: Experiments on using semantic distances between words in image caption retrieval. In: Proc. 19th ACM Int'l Conf. on Research and Development in Information Retrieval, Zurich (1996) 174–180
3. Chen, Y., Wang, J.Z., Krovetz, R.: CLUE: Cluster-based retrieval of images by unsupervised learning. IEEE Transactions on Image Processing **14** (2005) 1187–1201
4. Aizawa, A.: An approach to microscopic clustering of terms and documents. In: Proceedings of PRICAI 2002: Trends in Artificial Intelligence : 7th Pacific Rim International Conference on Artificial Intelligence. Volume 2417 of Lecture Notes in Computer Science. (2002) 404–413
5. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. **22** (2004) 179–214
6. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS, Alicante, Spain (2006) to appear
7. Inoue, M.: The remarkable search topic-finding task to share success stories of cross-language information retrieval. In: New Directions in Multilingual Information Access: A Workshop at SIGIR 2006, Seattle, USA (2006) 61–64