Searching by Similarity and Classifying Images on a Very Large Scale

Giuseppe Amato, Pasquale Savino
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche, ISTI-CNR
Pisa, Italy
{*giuseppe.amato,pasquale.savino*}*@isti.cnr.it*

***Abstract*—In the demonstration we will show a system for searching by similarity and automatically classifying images in a very large dataset. The demonstrated techniques are based on the use of the MI-File (Metric Inverted File) as the access method for executing similarity search efficiently. The MI-File is an access methods based on inverted files that relies on a space transformation that use the notion of perspective to decide about the similarity between two objects. More specifically, if two objects are close one to each other, also the view of the space from their position is similar. Leveraging on this space transformation, it is possible to use inverted file to execute approximate similarity search. In order to test the scalability of this access method, we inserted 106 millions images from the CoPhIR dataset and we created an on-line search engine that allows everybody to search in this dataset. In addition we also used this access methods to perform automatic classification on this very large image dataset. More specifically, we reformulated the classification problem, as resulting from the use of SVM with RBF kernel, as a complex approximate similarity search problem. In such a way, instead of comparing every single image against the classifier, the best images belonging to a class are directly obtained as the result of a complex approximate similarity search query.**

***Keywords*-similarity search; image content based retrieval; image classification**

I. Introduction

Large scale image content based retrieval is becoming realistic thanks to the recent advances on techniques for building access methods for scalable and efficient similarity search.

We will demonstrate two applications that respectively perform *image content based retrieval* and *image content classification* on the CoPhIR dataset [1] created in the SAPIR project [2].

CoPhIR consists of 106 millions images, taken from Flickr [3], described by MPEG-7 [4] visual descriptors. Content based retrieval can be performed by using similarity functions between the visual descriptors associated with the images.

II. Large Scale Image Search

The image search engine application is a web application that can be freely accessed through any internet connection. The home page of the application, as shown in Figure 1, allows the user to choose an image to be used as a query. The user, optionally, can also choose some parameters to decide a tradeoff between efficiency of the system and quality of the search results. These parameters will be briefly described in the next section.

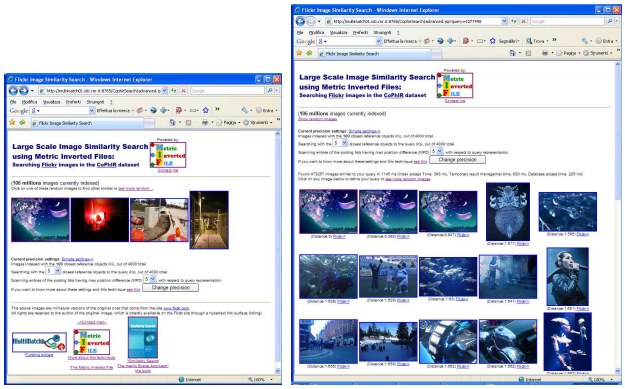

Figure 1. The image search engine: choosing an image query and seeing the result.

Once the user clicks on the query image, the system shows the 40 images that are deemed to be the most visually similar to the query. From the result page the user can choose another image to perform another query or go to the Flickr page where the original image can be accessed.

The application runs on a *single server* having 2 Xeon dual core CPUs running at 3.0 GHz and 16 Gb of main memory. The access method used to support efficient and scalable similarity search is the MI-File [5] (Metric Inverted File), which is briefly discussed in the following.

Tests with 10 really simultaneous users (simulated with a process consisting of 10 parallel threads each submitting a random query to the system) returned a throughput of 0.5 query per second executed.

The image search engine can be accessed at http://multimatch01.isti.cnr.it:8765/CophirSearch/

A. Scalable access method

The MI-File access method, used to support efficient and scalable similarity search, is a disk maintained index based on inverted files. More details can be found in [5]. It relies on a space transformation that uses the notion of perspective to decide about the similarity between two objects. More

specifically, if two objects are close one to each other, also the view of the space from their position is similar. To realize this idea, the MI-File chooses a set of reference objects RO from the dataset to be indexed. When an object o has to be inserted in the index, we first search the k_i ($k_i \leq \#RO$, but typically $k_i \ll \#RO$) objects of RO closest to o. The ordered sequence of the k_i reference objects $(ro^o_1, \ldots, ro^o_{k_i})$ is the representation of o in the transformed space. A similar space transformation was also proposed in [6], [7]. This representation can be conveniently maintained in an inverted file if we assume that each $ro \in RO$ is the entry of a posting list of the inverted file. The posting list associated with an entry $ro \in RO$ is a list of pairs (o, i), such that ro is the $i-th$ closest object to o among those in RO.

Suppose the user chooses q as a query image. In order to execute the query, first the $k_s (k_s \leq k_i)$ closest reference objects to q are retrieved. Then, the retrieved k_s reference objects are used to select the posting lists to be scanned, in order incrementally compute the *Induced Spearman Footrule Distance* distance [8] between the query representation and the image in the index. The query execution can be optimized asking the system just to scan the most promising portion of the posting lists. The value of k_s and the fraction of the posting lists to be accessed can be chosen by the user. More details on the MI-File technique can be found in [5].

III. Large Scale Image Classification

In addition to pure image similarity search, we will also demonstrate a technique of image classification that can be efficiently applied to a large scale dataset.

We defined a number of classifiers for various types of content (Church, Temple, painting, etc.) using Support Vector Machines (SVM) [9] with Radial Basis Function (RBF) kernel. We reformulated the decision function of the obtained classifiers as a complex similarity search problem by combining the support vectors obtained at the end of the learning process. Specifically, each positive support vector is used to perform a similarity search on the entire image dataset. Similarity searches are processed by the MI-FIle. The union of the results forms the candidate set to be classified using the decision function. In such a way, instead of applying the classifier to each of the 106 millions images sequentially, we apply the original decision function to the images in the candidate set in a reasonable time.

This technique allows, for instance, to easily refine a classifier or to use a new classifier on an existing dataset, given that the classifier does not need to be reapplied to every single image of the dataset: when a refined or new classifier is available, a complex similarity query is generated to obtain a set of candidates for the class.

The screen-shoot of the image classification web application is shown in Figure 2. The user can choose a class and can ask the system to show the pre-recomputed class or to reevaluate the classifier. Recomputing the class, using the MI-File to execute the corresponding complex similarity search, takes about four minutes. This is some orders of magnitude less than the 66 hours required to directly evaluate the classifier with each image in the dataset.

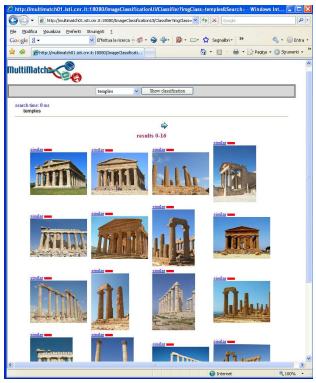

Figure 2. Image classification application.

The classifier can be accessed at http://multimatch01.isti.cnr.it:18080/ImageClassificationUI/?rate

References

[1] "CoPhIR: Content-based Photo Image Retrieval Test-Collection," Project Web Site, 2009, http://cophir.isti.cnr.it/.

[2] "SAPIR: Search In Audio Visual Content Using Peer-to-peer IR," Project Web Site, 2009, http://www.sapir.eu/.

[3] Http://www.flickr.com/.

[4] "Mpeg-7," ISO/IEC JTC1/SC29/WG11N6828, October 2004, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[5] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in *InfoScale '08: Proceedings of the 3rd international conference on Scalable information systems*. ICST, 2008, pp. 1–10.

[6] E. Chavez, K. Figueroa, and G. Navarro, "Proximity searching in high dimensional spaces with a proximity preserving order," *Lecture Notes in Computer Science*, vol. 3789, p. 405, 2005.

[7] E. Gonzalez, K. Figueroa, and G. Navarro, "Effective Proximity Retrieval by Ordering Permutations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1647–1658, 2008.

[8] P. Diaconis, *Group Representations in Probability and Statistics*, ser. IMS Lecture Notes - Monograph Series. Institute of Mathematical Statistics, Hawyard Ca, 1988, vol. 11.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0521780195