VALIDATION OF REMOTE SENSING CONTENT-BASED INFORMATION RETRIEVAL (RS-CBIR) SYSTEMS UPON SCARCE DATA

Miguel A. Veganzones, Manuel Graña

Grupo de Inteligencia Computacional, Universidad del Pais Vasco (UPV/EHU), Spain

ABSTRACT

Validation of Remote Sensing Content-Based Information Retrieval (RS-CBIR) systems requires innovative strategies to overcome the scarcity of labeled data. CBIR systems validation by means of precision/recall measures based on either, user's feedback or a-priori known categories, are hard to apply to RS-CBIR systems. **The main handicaps are the lack of ground truth knowledge (categories) and the user difficulties to evaluate the system's response images to provide a positive/negative feedback**. We propose to apply a data-driven (unsupervised) quality assessment strategy analogous to the DAMA strategy applied for the validation of classification methods used in thematic mapping. The strategy is intended for quality assessment when little or no ground truth is available. **The proposed strategy deals with the RS-CBIR validation problem by giving a quantitative and qualitative evidence of the relative (subjective) quality of RS-CBIR systems without a-priori knowledge**.

CBIR SYSTEMS

A CBIR system model is a tuple $M=\langle D,\phi,d,\psi\rangle$ where $D=\{x_i\}_{i=1}^n$ is a dataset with n images, $\phi(x)$ is a feature extraction process which maps any image x onto a feature space Φ; $d:\phi\times\phi\rightarrow\mathbb{R}^+$ is a distance function measuring the dissimilarity between two images defined on their features; and, ψ is an optional retrieval feedback process, which allows the user to provide a feedback to the CBIR system to improve the data search process.

CBIR systems can be modeled as a two-class problem where images are classified as belonging to the relevant class, $C_R(q)$, or to the irrelevant class, $C_I(q)$, relative to a query q. The ideal response $f_M(q)$ of a CBIR system M to a query q is given by

$$f_M(q)=\{x_{i_1},\ldots,x_{i_k},x_{i_{k+1}},\ldots,x_{i_n}\} \qquad (1)$$

where images $\{x_{i_1},\ldots,x_{i_k}\}$ belong to the relevant class $C_R(q)$, and images $\{x_{i_1},\ldots,x_{i_k}\}$ belong to the irrelevant class $C_I(q)$.

Thus, the expected response of an ideal CBIR system, $f_M(q)$, to a potential user's query q, can be seen as a relevant/irrelevant mapping of the dataset D in an unknown feature space, Ω. We name this ideal mapping the *potential search map*.

The performance of a CBIR system is related to its capacity to elaborate a mapping equivalent to the potential search map of any given query.

The two most frequently used evaluation measures are *precision* and *recall*. If we denote *T* the set of returned images and *R* the set of all the images relevant to the query, then

$$precision=\frac{|T\cap R|}{|T|} \qquad recall=\frac{|T\cap R|}{|R|} \qquad (2)$$

Results are usually summarized as precision-recall or precision-scope curves. Using the concepts defined above, for a given query q, the set R of images relevant to the query identifies with the relevant class specified by the potential search map of the query: $R=C_R(q)$.

PROPOSED VALIDATION STRATEGY

We propose to perform clustering processes over a dataset D to discover data inherent equivalence relation structures, and to use this inherent structures to simulate potential user queries.

Each clustering can be used to model potential search maps of a family of queries Q. Then, the simulated potential search maps are used to provide precision and recall measures, that show the RS-CBIR system capacity to solve the family of queries Q on a dataset.

The procedure is as follows:

1. Perform a clustering over a dataset D. The clustering process defines a mapping $X^*=\{x_1^*,\ldots,x_n^*\}$ where x_i^* indicates the identity of cluster c_k, $k=1,\ldots,C$, which image i belongs to, where C is the number of clusters found in the clustering process. This clustering represents a family of queries, $Q=\{q_1,\ldots,q_n\}$, where each query q_j is given by the sample image x_j, $j=1,\ldots,n$.
2. Given a RS-CBIR system M, calculate its response to each of the queries $q_j\in Q$, $f_M(q_j)=\{x_{i_1},\ldots,x_{i_n}\}_{j=1}^n$, which can be represented as a matrix $A=\{a_{ij}\}, i,j=1,\ldots,n$, so a_{ij} indicates the *i*-th most similar image to the query q_j.
3. Being *s* the query scope, the set of returned images T_j and the set of all the relevant images R_j for a query q_j are:
$$T_j=\{U_{i=1}^s a_{ij}\} \quad R_j=\{U_{x_i^*} x_i;q_j\in c_k\} \qquad (3)$$
4. Now, the precision and recall measures for the query q_j can be calculated by substituting (3) in equations (2). The average of the precision and recall measures estimated by all the queries $q_j\in Q$ is a quality assessment of the RS-CBIR system *M* response respect to the family of queries *Q* on *D*.

CONCLUSIONS AND FUTHER WORK

There is a strong need of innovative strategies to validate RS-CBIR systems that could successfully overcome the lack of ground truth data. We were inspired by the DAMA strategy to work in quality assessment techniques for RS-CBIR systems validation. We have formalized an strategy to assess the quality of RS-CBIR systems when little or no ground truth data is available, modeling potential users queries by clustering processes, and providing a mechanism to define the set of relevant images for a given query, in order to provide precision and recall quality measures. Futher work will involve applying the proposed validation strategy in developed RS-CBIR systems.

Contact: {miguelangel.veganzones,manuel.grana}@ehu.es
Http://www.ehu.es/computationalintelligence