Astronomical Data Analysis Software and Systems VII
ASP Conference Series, Vol. 145, 1998
R. Albrecht, R. N. Hook and H. A. Bushouse, eds.

Querying by Example Astronomical Archives

F. Pasian and R. Smareglia

Osservatorio Astronomico di Trieste, 34131 Trieste, Italy
Email: pasian@ts.astro.it

Abstract. In this paper, a project aiming to allow access to data archives by means of a query-by-example mechanism is described. The basic user query would simply be the submission of an input image, either 1-D or 2-D; the system should allow retrieval of archived images similar to an input one on the basis of some "resemblance parameter".

1. Introduction

With the exponential expansion of information systems, accessing data archives plays a role which is becoming increasingly important for the dissemination of scientific knowledge at all levels. The need is felt to make access to data archives simpler for the community of users at large. Queries on the data, regardless of their complexity, should be easier to specify, more intuitive, simplified with respect to those currently in use. In other words, users should be able to express queries in an intuitive way, without knowing the detailed physical structure and naming of data. Content-based information retrieval is receiving increasing attention from scientists, e.g., see Ardizzone et al. (1996), and Lesteven et al. (1996).

This project aims to allow access to data archives by means of a query-by-example mechanism: the user should be put in a position to provide a "user object", a template of what he/she wishes to retrieve from the archive. The system managing the archive should provide the user with archived data resembling the template on the basis of some desired characteristic, and with some "resemblance parameter". Data understanding (classification and recognition of descriptive features) is deemed to be an essential step of the query-by-example mechanism.

2. Project Objectives

The objective of this project is to integrate database, mass storage, classification - feature recognition, and networking aspects into a unique system allowing a "query-by-example" approach to the retrieval of data from remotely-accessible large image archives exploiting the algorithms up to their performance limits.

The basic idea, from the user's point of view, is to be able to submit a query to a remote archive in the form of an image (2-D or 1-D), telling the system: *"get me all the images/plots in the archive looking like this one, or having this*

specific feature". The basic user query would therefore simply be the submission of an image, either previously extracted from the very same archive, or owned by the user, or built via a modelling software. This operation must be feasible while connected remotely to the archive.

Several problems have been already identified. In the "global" approach to the query-by-example mechanism, the various characteristics of images have the same weight. Difficulties here include coping with different image resolutions, being able to search for images having a specific feature in them by just submitting a small image template containing the feature itself, building models compatible with the archived images, *etc.* A more sophisticated approach would be to add a level of interaction with the user, allowing him/her to know which are the features of the images the system is able to recognize, and allow a choice of the features to be searched for.

3. QUBE - Overall Features

A system called QUBE is currently being designed to support the QUery-By-Example paradigm in accessing astronomical data archives. A preliminary list of features is as follows:

- Network-based interface, for backward-compatibility with already existing image archives: standard queries on metadata (data descriptions) should always be possible through the known interface.
- Extensions to standard SQL, to allow handling of image templates by a relational database: in principle, specifying a query in this extended SQL should always be possible for an expert user. As a first order approximation, SQL extensions can be handled by an interpreter.
- Ingestion of "user objects" in the system in different formats, with priority given to astronomy-specific and commercial standards (FITS, but also GIF, JPEG, etc.).
- "Global" approach to query-by-example based on classification methods such as artificial neural networks: unsupervised in the general case, supervised for specific applications.
- Feature recognition algorithms may be used, able to give different weights to different characteristics of the "user object" (the template being compared with the archived data).
- Transmission of data (both "user objects" and retrieved data) via the standard TCP/IP mechanisms.
- An updating capability should be available: in the case classification is required on an image subset (e.g., wavelength range in a spectral archive) the system should be able to re-compute classification parameters for all relevant image subsets in the archive.
- A feedback mechanism should be built in the system, to allow detection and correction of misclassified data.

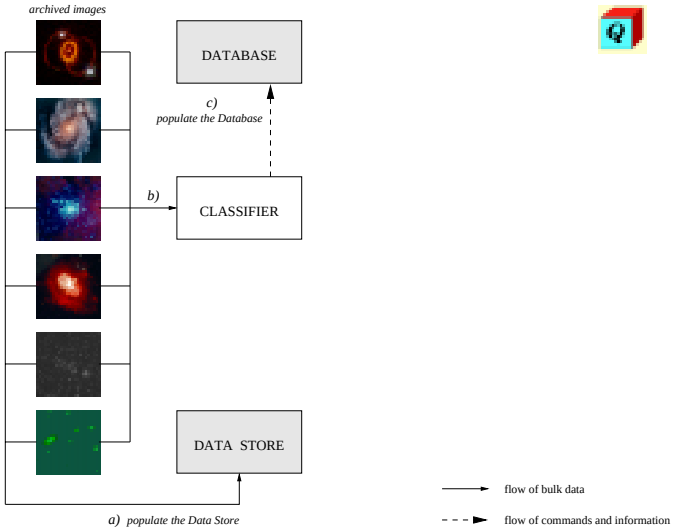

Figure 1. Operational scenario for the data ingest phase of the QUBE system

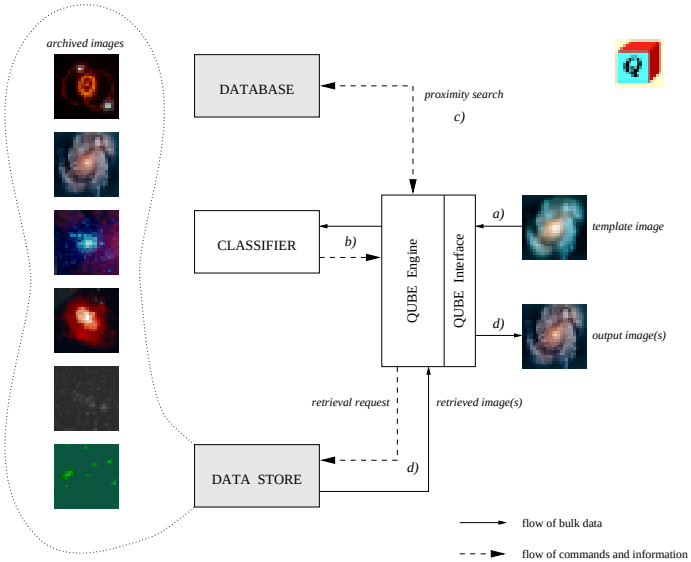

Figure 2. Operational scenario for the data retrieval phase of the QUBE system

4. QUBE - Structure and Operational scenario

The QUBE system is composed of:

- the Archive (a Data Store containing the images, using optical media and a jukebox system; a Database containing their descriptions, managed by a relational DBMS);
- a Classifier (parametric or based on Artificial Neural Networks);
- the QUBE Search Engine;
- the QUBE User Interface.

The following operational scenario is envisaged, and is represented graphically in Figures 1 and 2, respectively:

Data ingest phase : When ingested in the Data Store (step *a*), the images are also analyzed by a Classifier, either parametric or ANN-based, as in Pasian et al. (1997) (step *b*). The result of the classification phase is ingested in the Database (step *c*), together with the image parameters, e.g., extracted from the FITS headers.

Data retrieval phase : A template image is given by the user as input to the system (step *a*). The QUBE Search Engine re-scales it according to the resolution of the archived images, possibly subsets it, and submits it to the Classifier, obtaining the classification parameters (step *b*). A proximity search is then performed on the Database (step *c*); the image(s) satisfying the desired conditions are identified, retrieved from the Data Store and fed to the user as output (step *d*).

More than one classifier may be envisaged to make the mechanism more flexible: in this case, more than one table containing the classification parameters for the archived images will be stored in the Database.

Acknowledgments. The authors are grateful to H.M.Adorf, O.Yu.Malkov, J.D.Ponz and M.Pucillo for having discussions about the concepts of querying image archives by example. The images used for the figures in this paper are thumbnails of Hubble Space Telescope PR images.

References

Ardizzone, E., Di Gesù, V., & Maccarone, M. C. 1996, in Strategies and Techniques of Information in Astronomy, Vistas in Astronomy, 40, 401

Lesteven, S., Poinçot, P., & Murtagh, F. 1996, in Strategies and Techniques of Information in Astronomy, Vistas in Astronomy, 40, 395

Pasian, F., Smareglia, R., Hantzios, P., Dapergolas, A., & Bellas-Velidis, I. 1997, in: Wide-Field Spectroscopy, E.Kontizas, M.Kontizas, D.H.Morgan, G. Vettolani eds., Kluwer Academic Publishers, 103