Astronomy and Computing 14 (2016) 23–34

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Indexing data cubes for content-based searches in radio astronomy

M. Araya, G. Candia, R. Gregorio, M. Mendoza*, M. Solar

Department of Informatics, Universidad Técnica Federico Santa María, Vicuña Mackenna 3939, San Joaquín, Santiago, Chile

ARTICLE INFO

Article history:
Received 19 June 2015
Accepted 1 January 2016
Available online 19 January 2016

Keywords:
Data cubes
Virtual observatories
Content-based search
Astroinformatics

ABSTRACT

Methods for observing space have changed profoundly in the past few decades. The methods needed to detect and record astronomical objects have shifted from conventional observations in the optical range to more sophisticated methods which permit the detection of not only the shape of an object but also the velocity and frequency of emissions in the millimeter-scale wavelength range and the chemical substances from which they originate. The consolidation of radio astronomy through a range of global-scale projects such as the Very Long Baseline Array (VLBA) and the Atacama Large Millimeter/submillimeter Array (ALMA) reinforces the need to develop better methods of data processing that can automatically detect regions of interest (ROIs) within data cubes (position–position–velocity), index them and facilitate subsequent searches via methods based on queries using spatial coordinates and/or velocity ranges. In this article, we present the development of an automatic system for indexing ROIs in data cubes that is capable of automatically detecting and recording ROIs while reducing the necessary storage space. The system is able to process data cubes containing megabytes of data in fractions of a second without human supervision, thus allowing it to be incorporated into a production line for displaying objects in a virtual observatory. We conducted a set of comprehensive experiments to illustrate how our system works. As a result, an index of 3% of the input size was stored in a spatial database, representing a compression ratio equal to 33:1 over an input of 20.875 GB, achieving an index of 773 MB approximately. On the other hand, a single query can be evaluated over our system in a fraction of second, showing that the indexing step works as a shock-absorber of the computational time involved in data cube processing. The system forms part of the Chilean Virtual Observatory (ChiVO), an initiative which belongs to the International Virtual Observatory Alliance (IVOA) that seeks to provide the capability of content-based searches on data cubes to the astronomical community.

1. Introduction

Modern astronomy is characterized by the accumulation of novel methods for observing astronomical objects. Many of these have shifted toward methods based on recording emissions on millimeter-scale wavelength ranges, observations that enable the determination of the chemical compositions of detected astronomical objects. The interest in exploring the existence of substances such as CO in galactic or extra-galactic observations has spurred the development of numerous projects, such as the Very Long Baseline Array (VLBA) (Very Long Baseline Array, VLBA, 2016) and the Atacama Large Millimeter/submillimeter Array (ALMA) (Atacama Large mm/submm Array, 2016), which are capable of recording the frequencies and velocities of emission lines from objects at astronomical distances producing a growing data archive for astronomy research purposes. Furthermore, it is projected that when the Square Kilometer Array (SKA) (Square Kilometre Array, SKA, 2016) begins operation in 2020 more than 60 PB of archived data will be accessible to astronomers (Berriman and Groom, 2011). This will produce a high amount of daily data traffic, creating a need for access to data processing methods that are capable of contending with such large data sets. For example, it is estimated that when it is running at full capacity, ALMA will generate more than 750 GB of data every day (approximately 250 TB a year) (Atacama Large mm/submm Array, 2016). Therefore, the astronomical community will require the use of a high-speed data transmission system to archive the data of interest and analyze them to extract information relevant to their needs. Because of the enormous volume of data that will be generated, it will be impractical for analytical procedures to be performed on the entire data set. The need for more and better automatic detection, indexing, recording, and cataloging methods is thus a key factor in the continued growth of astronomy in the 21st century.

* Correspondence to: Vicuña Mackenna Avenue 3939, San Joaquín, Santiago, PO 8940000, Chile.

E-mail address: marcelo.mendoza@usm.cl (M. Mendoza).

http://dx.doi.org/10.1016/j.ascom.2016.01.002
2213-1337/© 2016 Elsevier B.V. All rights reserved.

The ultimate objective of developing automatic detection, indexing, recording, and cataloging systems is to provide integrated systems with virtual search software, also known as virtual observatories (Araya et al., 2015). Virtual observatory development initiatives are coordinated through the International Virtual Observatory Alliance (IVOA) (International Virtual Observatory Alliance, 2016), which catalogs virtual observatories around the world that provide access to and search methods for extensive collections of astronomical objects. In particular, the emergence of the Chilean Virtual Observatory (ChiVO) (Chilean Virtual Observatory, 2015) and its incorporation into the IVOA has spurred the development of a system for detecting and recording regions of interest (ROIs) in radio astronomy data, which allows for content-based searches as part of ChiVO.

In this article, we present the methods and techniques used to develop the data cube indexing system for content-based searches as part of ChiVO. The indexing system was designed under efficiency constraints and therefore incorporates computationally lightweight processes that are capable of single-pass data processing – thus facilitating its implementation – and the handling of large amounts of data while significantly reducing the cost needed for content-based searches. Our goal is to build an effective indexing system, trying to absorb computational costs at the indexing step reducing the time involved in data recovery. We will show in our experiments that our system is capable to create an index at 33:1 compression ratio. The index helps us to process spatial queries in fraction of seconds, showing that the indexing step works as a shock-absorber of the computational time involved in data processing. Thus, the key factor of our system is the process of creation of the index, where a number of design decision are taken to address the tradeoff between quality of approximation and computational time involved in data processing.

The article documents, step by step, the methods used for data processing, the strategies used for signal/noise processing, the techniques used for spectrogram processing and for obtaining summary spectrograms that enable the determination of velocity ranges of interest, the methods used for stacking data slices within ranges of interest over which morphological structuration processes are applied, and the methods that assist in identifying the localization parameters for objects.

The contributions of this article include providing a detailed description of the data indexing system to serve as a potential model for future work in the field of data cube indexing. The article also presents solutions to the problem of large-scale data processing through the use of low-cost computational operations, thus addressing the requirement for online processing under high-demand conditions. This article is specifically directed to the community of astroinformatics software developers, but the concepts presented herein are also generally useful for all types of software development with a need to solve problems of large-scale data processing and indexing, particularly for multiway data.

This article is organized into the following sections. Section 2 presents a review of related works. Section 3 discusses the concepts of data cubes, morphological structuration, and shape detection. Section 4 presents the general architecture of the spectrographic cube indexing system. Sections 5 and 6 discuss the system components related to spectrographic processing and ROI detection and indexing, respectively. Section 7 presents a comprehensive set of experiments to validate our proposal. Conclusions are presented in Section 8.

2. Related work

The field of radio astronomy software development has been quite active in the past few decades. Current software has been developed primarily for the manipulation, visualization, and post-processing analysis of data; this is significantly different from our system, which was designed for the online indexing and recording of data cubes. Consequently, the former systems, designed for radio astronomer end users, are restricted not by the computational costs of the methods used but rather by the effectiveness of the analysis methods. Our system, by contrast, is designed to be implemented in a virtual observatory (VO) and, therefore, to provide indexing services to the VO, which itself provides the data to radio astronomers.

The majority of developed software provides functionalities for performing spectrographic processing within proprietary software such as IDL. For example, GBTIDL (National Radio Astronomy Observatory, GBTIDL, 2005) is an interactive package for the reduction and analysis of spectral line data extracted from data cubes. The package consists of calibration and analysis procedures for cubes from the Green Bank Telescope (National Radio Astronomy Observatory, Green Bank Site, GBT, 2016) using the observatory's proprietary format, GBT SDFITS. Rosolowsky and Leroy developed CPROPS (Rosolowsky and Leroy, 2006) with the primary purpose of characterizing molecular clouds. The software uses CLUMPFIND, an algorithm developed by Williams et al. (1994), which locates optima in 3D data. MIR (Qi, 2014) is a software application that processes cubes obtained from the Smithsonian Submillimeter Array (SMA) (Smithsonian Astrophysical Observatory, SMA, 2005). The primary purpose of this software is to provide routines for interactive data analysis and visualization. Spectroscopy Made Easy (SME) (Valenti and Piskunov, 1996) was developed to process stellar spectra by fitting a synthesized spectrum to the data, from which relevant spectroscopic parameters could be inferred. Valenti and Fischer (Valenti and Piskunov, 1996) demonstrated that SME could be used to obtain spectroscopic parameters from thousands of stars. All of the above programs were developed for IDL.

Spectrographic analysis software has also been developed with even more specific purposes in astronomy. For example, Kochukhov (Kochukhov et al., 2010) developed software to calculate the intensity of stellar spectra for a given atmospheric model under the assumption of thermodynamic equilibrium. SprexTool (Cushing et al., 2004) is a spectral line extraction package developed specifically for infrared observations. DUCHAMP (Whiting, 2012) was designed to find sources in three-dimensional spectral-line data cubes, being designed for H1 analysis but being possible to extend its application to other molecules. Recently, Serra et al. introduced SoFiA (Serra et al., 2015), a source finder for 3D spectral line data which comprises source-finding algorithms for H1 surveys. These tools are useful for spectral line analysis, working with catalogs in a tandem for source finding, illustrating how relevant is the design of effective analysis tools for data cube processing. We consider this kind of tools as a complement of our system, which can be considered as a pre-processing system for data cube processing in a large scale storage system.

For molecular cloud analysis a variety of clump-finding algorithms has been proposed. The CUPID software package (Berry, 2007) includes a modern compilation of them. The CLUMPFIND (Williams et al., 1994) algorithm, which was designed to analyze molecular clouds in radio astronomy, it operates into three dimensions (position–position–velocity), seeking out local emission peaks and following them to lower intensity levels. The result is the detection of structural units, or clumps, in which emission is concentrated. CLUMPFIND was originally part of the data reduction package MIRIAD (Sault et al., 1995), which is used by the BIMA interferometer (Berkeley Illinois Maryland Association, 2004). Despite the popularity of CLUMPFIND in molecular cloud analysis, the technique is not free from criticism. From the astrophysical perspective, the generated clump-list of crowded regions