El grupo de Ciencia con Observatorios Virtuales del INAOE

R. Terlevich[1], A. López López[1] y E. Terlevich[1] Junio 2003

1 - INOAE, Instituto Nacional de Astrofísica Óptica y Electrónica

Abstract. Se acerca un tsunami de datos observacionales de gran calidad para los cuales la comunidad astronomica no está aún preparada.

Esto es debido en gran parte a que en los últimos años, se ha presenciado la maduración de la Cosmología observacional como rama dominante de la Astrofísica, combinado con la puesta en marcha de grandes telescopios y de un interés rápidamente creciente en grandes proyectos observacionales destinados a comprender la estructura y evolución del Universo. México, con el GTM, será un contribuyente clave en esta área

Hoy en día, la mayoría de los proyectos observacionales de más envergadura, involucran entre otras cosas el recolectar y archivar enormes conjuntos de datos. Las dimensiones de estos conjuntos superan ampliamente las capacidades actuales de manejo y análisis de la comunidad astronómica.

El objetivo de nuestro grupo es proveer la infraestructura para la formación de grupos interdisciplinarios con preparación y experiencia en computación, astrofísica, estadística, dinámica y visualización, como para poder aprovechar el flujo de datos (que se anticipa masivo) de experimentos presentes y futuros relacionados con Cosmología, realizados tanto desde tierra como desde el espacio. El objetivo último es el desarrollo de un Observatorio Virtual que permitirá no sólo acceso a las bases de datos sino además la infraestructura y programas necesarios para realizar las investigaciones asociadas.

1 Por qué un grupo de Ciencia con Observatorios Virtuales?

Ha habido en los últimos años un incremento colosal en la cantidad disponible de datos para la investigación astronómica, debido en parte a la gran cantidad de facilidades nuevas, en parte a avances tecnológicos en sensores y dispositivos de almacenaje, y finalmente, a la creciente automatización de la adquisición y procesamiento de datos.

Como consecuencia de todo esto, en los próximos años, los astrónomos serán capaces de generar datos calibrados más rápidamente que los que puedan procesar y analizar a conciencia.

Nuevos relevamientos extensos como el 2dF, SDSS, 2MASS, VIRMOS y DEEP2, están revolucionando la manera que se hace astronomía, proporcionando cantidades enormes de datos de alta calidad, además por supuesto de los datos ya existentes. A esta marea gigante de datos ópticos, obtenidos con

telescopios de tamaño mediano, debemos agregar las vastas bases de datos que serán producidas por la nueva generación de grandes telescopios de tierra y del espacio, en rayos-X, óptico, infrarrojo y radio (mm).

La naturaleza de estas bases de datos (amplia cobertura en longitudes de onda), constituirá un desafío sin precedentes para la unificación de observaciones ópticas, infrarrojas, milimétricas y de rayos-X, obtenidas desde observatorios espaciales y de tierra, los que incluyen también radio-telescopios como el GTM (Gran Telescopio Milimétrico, que está siendo construído actualmente por el INAOE y la Universidad de Massachusetts en el Cerro La Negra, México, a 4600 m de altura).

Es imposible exagerar la importancia y complejidad de este desafío. En los próximos cinco años, los datos disponibles en bases de datos crecerán hasta alcanzar un nivel de decenas de miles de parámetros para centenas de millones de objetos astronómicos, o sea, TERADATASETS. Esta complejidad se verá aumentada por la existencia de errores de medición, desviaciones y tendencias en los datos, y sobre todo la dificultad más importante que es la falta de homologación a través de las distintas bases de datos, cada una de las cuales tendrá no solo su propio conjunto de parámetros pero además su propio software de acceso lo que dificultará extremadamente la correlación cruzada entre las mismas.

Para poder avanzar, bajo el peso continuamente creciente de la cantidad de datos, se requieren aplicaciones que, en forma inteligente y automática, manejen, formulen preguntas, visualicen y analicen el espacio completo y variado de las grandes bases de datos, o sea un Observatorio Virtual (OV). Esto es más importante aún para países que, como México, no cuentan con acceso a las grandes facilidades observacionales internacionales

INAOE está particularmente bien situado para jugar un papel importante en este nuevo campo. El tamaño, calidad y proximidad física de los Departamentos de Astrofísica y Computación es excepcional entre los Departamentos de Astronomía a nivel mundial y permite el desarrollo rápido y eficiente de grupos específicos interdisciplinarios, como el que presentamos aquí. La colaboración existente con el Instituto de Astronomía de la Universidad Cambridge y con el Departamento de Astronomía del University College de Londres ya nos brinda, entre otras cosas, acceso a las muchas bases de datos británicas y europeas.

En cuanto a las oportunidades para enseñanza y entrenamiento, la combinación de computación, estadística, visualización y rigor científico, resulta en una sinergía poderosa, con amplias aplicaciones al quehacer de la sociedad, y no sólo a la investigación pura. Desde economía y bancos hasta medicina, hay una amplia variedad de aplicaciones que requieren de expertos en este nuevo campo. Nuestro grupo estará óptimamente situado para distribuir este conocimiento a toda la sociedad mexicana.

Un objetivo importante a mediano plazo es el de servir de base para un futuro Observatorio Virtual Mexicano (OVM) que permita a toda la comunidad astronómica nacional no solo acceso a los datos sino que además integre toda la infraestructura necesaria para realizar las investigaciones al mismo nivel que otros Observatorios Virtuales. El plan es además integrar el OVM al Interna-

tional Virtual Observatory Alliance (IVOA) con el objeto de facilitar la coordinación y alcanzar con el máximo éxito y en el menor tiempo los objetivos del grupo.

Dentro del IVOA, cada esfuerzo nacional tendrá sus objetivos particulares definidos en un conjunto de objetivos cientificos, intereses tecnológicos y sus propias medidas de éxito. Esta diversidad es importante para cubrir el enorme espacio de interés de un OV, pero también debe haber elementos del esfuerzo internacional que es conveniente coordinar para que el IVO sea una realidad y además evitar duplicación de esfuerzos. Los elementos más comunes para coordinar son los estándares de los datos y las interfaces para superar la inherente imposibilidad de intercambiar datos entre los distintos OV. Otros elementos comunes pueden ser paquetes de software, códigos, bibliotecas de programas y herramientas de desarrollo.

Consideramos de gran importancia que el futuro OVM se integre al IVOA.

Inicialmente, un modesto presupuesto será necesario para educación y entrenamiento más un programa de intercambio y de visitantes, que permitirá establecer y mantener colaboraciones con otras instituciones.

2 Objetivos

Hemos elegido un plan de aprendizaje que, a partir de objetivos relativamente modestos, apunta a resultados ambiciosos teniendo muy en cuenta la produccion de resultados intermedios en la forma de publicaciones de gran calidad y el entrenamiento de estudiantes de posgrado.

Estamos interesados en el desarrollo y comparación de diversos métodos estadísticos que permitan analizar TERADATASETS múltiples.

Nuestros objetivos principales son varios e incluyen:

- Desarrollo de una interfaz entre el usuario y las diferentes bases de datos, basada en análisis remoto.
- Definir los parámetros que caracterizan rasgos prominentes y dividir las bases de datos en muestras y sub-muestras que se aproximen a unos 1000 gigabytes cada una.
- Desarrollo de herramientas de visualización.
- Clasificación sin supervisión y determinación de los distintos números de clases de objetos presentes en los datos.
- Búsqueda de objetos inusuales, incluyendo la habilidad de detectar nuevas clases de objetos.
- Búsqueda de regularidades o correlaciones en los datos.
- Apoyo para análisis supervisado.
- En paralelo y utilizando técnicas similares se desarrollara el análisis de las bases de datos producto de simulaciones numericas. En particular analisis de resultados de los nuevos modelos de síntesis de poblaciones estelares de alta resolución espectral.

Los resultados de estos análisis iniciales deberán, al mismo tiempo, proveer nuevas bases de datos para trabajo detallado de seguimiento. Esto será muy relevante en particular, para temas en los que estamos interesados, tales como la evolución de las propiedades de galaxias con líneas de emisión y su potencial uso para relevamientos de la evolución y geometría del universo; el estudio de las poblaciones estelares en galaxias normales y activas; la medición de la abundacia primordial de Helio.

3 Pasos a seguir

3.1 Bases de datos observacionales

Nuestro plan inicial está basado en el análisis local de los datos públicos y elimina por lo tanto la necesidad de transferir decenas de gigabytes de datos a través de una red que se encuentra a menudo sobrecargada. Este es el real 'cuello de botella' en muchas áreas de física computacional – a menudo se requieren todos, o la mayoría de los datos, en la memoria para efectuar el más simple de los análisis. El software que se desarrolle, podrá correr en sistemas baratos pero con gran capacidad de almacenamiento. Para comenzar este proceso, hemos elegido el Sloane Digital Sky Survey (SDSS), que representa quizás el más ambicioso relevamiento espectroscópico y de imágenes del Universo cercano, pero que además incluye también objetos lejanos y luminosos.

Simultáneamente, podremos comenzar a explorar técnicas y métodos para analizar submuestras relativamente pequeñas de datos (10^4 objetos con alrededor de 100 parámetros cada uno). Una vez que tengamos las facilidades de acceso y consulta, podemos intentar aplicar las mismas técnicas a muestras más grandes. Más adelante en el proceso de aprendizaje (dada nuestra relativa falta de experiencia en esta área particular) podremos proceder a trabajar en temas de visualización y presentación. Al momento, los productos disponibles para los datos del SLOAN (SDSS) incluyen ya un catálogo, que permite las búsquedas, conteniendo los objetos detectados y las imágenes y parámetros o atributos espectrales que se les asocian, imágenes a tres colores en formato JPEG, imágenes de datos en formato FITS y espectros en formato tanto GIF como FITS. La primera liberación de los datos del SLOAN (EDR, por las iniciales en inglés) cubre unos 462 grados cuadrados de cielo. Nuestra aproximación al problema será usar el EDR del SDSS para nuestra primera curva de aprendizaje. El EDR incluye más de 55000 espectros de galaxias, quasares y estrellas

A comienzos de este año, se hizo pública la primera liberación de datos del SDSS (DR1) que incluye espectros de 186240 objetos cubriendo 1556 grados cuadrados del cielo. Estos corresponden a 134015 galaxias, 17705 quasares, 17623 estrellas, 9684 espectros de cielo, 4491 espectros de estrellas dominados por bandas moleculares (M o más tardías), 1738 espectros con clasificación desconocida y 984 quasares a corrimientos al rojo mayores que 2.3.

Una cantidad considerable de trabajo ya ha sido hecha por Ofer Lahav y sus colaboradores en el IoA, Cambridge (en Septiembre Ofer y su grupo se mudarán

al University College London, en Londres) que han analizado y clasificado el conjunto de espectros de galaxias del 2dF por el método de análisis de componentes principales (PCA) y también utilizando otros métodos (ver, e.g. la revisión al respecto en Lahav, 2001). El relevamiento 2dF fue luego dividido de manera de obtener agrupamientos de funciones de luminosidad de acuerdo a clases espectrales (Madgwick et al. 2001 y otro en preparación). Ya se ha hecho un primer intento de detectar objetos inusuales (Madgwick et al. 2002). Métodos de clasificación no supervisados han sido ya aplicados Olac Fuentes a muestras astronómicas.

Tanto para las bases de datos empiricas como las teoricas, planeamos usar un ataque a dos puntas para su análisis y estudio:

1 - La aproximación objetiva ("sin supervisar") donde se deja que los datos "hablen por sí mismos" (e.g. PCA).

2 - El análisis supervisado (motivado físicamente) por ejemplo, en el análisis de índices e intensidades de líneas tradicionales, como Hα, [OIII], Mg2, etc.

Esta combinación de métodos nos parece importante para obtener relaciones entre parámetros físicos que sean astrofísicamente interesantes. Estas nuevas relaciones constituirán la base para estudios subsiguientes de evolución de galaxias y efectos ambientales.

Los resultados del análisis multivariado del DR1, serán usados para seleccionar sub-muestras interesantes de galaxias con líneas de emisión para estudiar, entre otras cosas:

- Edades de galaxias starburst (con brotes de formación estelar) y HII
- Distribución de abundancias de los elementos químicos y su evolución al mirar hacia atrás en el tiempo.
- Su utilización como estimadores de distancia a altos corrimientos al rojo y la determinación de parámetros cosmológicos.

Los estudios de seguimiento serán llevados a cabo usando grandes facilidades y consistirán en observaciones espectroscópicas de alta relación señal a ruido con múltiples rendijas de un número moderado de objetos seleccionados de los relevamientos SDSS y 2dF. Dado que estos relevamientos están basados en observaciones con fibras en telescopios de 3-4 metros, la calidad de los datos se verá incrementada sustancialmente con espectroscopía de rendija en telescopios de 8 metros.

3.2 Bases de datos teóricas

Otro aspecto central de nuestras tareas es el analisís de las bases de datos producto de modelos de síntesis de poblaciones estelares. La colaboracion del INAOE con Padova ha producido los primeros modelos de síntesis de poblaciones estelares de muy alta resolucion espectral.

El análisis de estos modelos figura con una prioridad altísima en nuestro trabajo inmediato y aplicaremos técnicas supervisadas y no supervisadas similares a las utilizadas con el análisis de bases de datos empiricas.

4 Estimación de la necesidades básicas del Grupo

Dada la falta de tiempo para hacer una estimación detallada y debidamente justificada las necesidades aqui listadas son solo aproximadas.

4.1 Para la segunda mitad del 2003

- Un sistema Raid de discos de 1 Terabite. (9000 US dolares)
- Dos escritores de DVD rápidos.(HP 300xi Interno; 230 US dolares cada uno)
- Ampliacion del almacenamiento en las computadoras de los miembros del grupo (Maxtor 160Gby EIDE 285 US dolares cada uno)
- Impresora laser. (HP Laser Jet 4300 2500 US dolares)
- Espacio fisico de al menos 15 m^2 para ubicar los equipos del grupo y algunos de los estudiantes
- Dos viajes para el workshop de IVOA en Strasbourg (Francia) en Octubre del 2003. (Pasajes 1000 dolares; Estadia 6 dias a 100 dolares por dia. Total 3200 dolares)
- Incorporación de estudiantes de Licenciatura - RJT tiene fondos disponibles de su SNI III como para dos estudiantes adicionales para este proyecto.

4.2 Para el 2004

- Adquisicion de un server ultrarrapido (DELL PE 6600 4 procesadores de 1.9Ghz; 8 HDD SCSI de 150 GB; 16Gby SDRAM; Linux. 350.000 pesos Mexicanos)
- Un sistema Raid de discos de 2 Terabites. (15.000 dolares. Si se compra el server DELL con sus 8 discos, se podria evitar este segundo RAID)
- Impresora laser color (HP 3800 dolares)
- Ocho pantallas TFT de alta resolucion (585 dolares cada una)
- Tres tecnicos de programacion con perfil a especificar
- Ampliacion del espacio fisico a 30m^2
- Seis viajes a reuniones internacionales (9600 dolares)
- Tres visitas al INOAE de expertos nacionales (10000 pesos mexicanos)
- Tres visitas al INAOE de expertos extranacionales.(3200 dolares)

4.3 Para el 2005

- Preveer la necesidad de un sistema de computacion de alta performance (150,000 dolares)

Referencias

Fuentes, Olac, 2001 Experimental Astronomy, Vol. 12, No. 1, pp. 21-31
Lahav, O., 2001, Proceedings of MPA/MPE/ESO Conference "Mining the Sky", 2000, Garching, Germany, astro-ph/0012407
Madgwick, D.S., Hewett, P.C., Mortlock, D.J.,
Lahav, O., MNRAS, astro-ph/0203307 (Hobson et al. 2002, 335,377)
Madgwick, D.S. and the 2dF team, 2001, MNRAS, astro-ph/0107197
Melnick, J., Terlevich, R. and Terlevich, E., 2000, MNRAS,311,629
astro-ph/9911094 Kunth and Ostlin, 2000, A+A Review, 10
astro-ph/0208246 Djorgovski et al.
http://www.nvosdt.org

El grupo, hasta ahora (30/06/2003)

Roberto Terlevich (IP), INAOE ext.2307. email: rjt@inaoep.mx

Aurelio López López(IP), INAOE ext.8314, 8318. email: allopez@inaoep.mx

Jesús A. González Bernal, INAOE. ext.8303 email: jagonzalez@inaoep.mx

Jaime Muñoz Arteaga, INAOE. ext.8313 email: jaime@inaoep.mx

Olac Fuentes Chávez, INAOE. ext.8304 email: fuentes@inaoep.mx

Elena Terlevich, INAOE ext.1314. email: eterlevi@inaoep.mx

Itziar Aretxaga, INAOE ext.2316. email: itziar@inaoep.mx

Diego Malquori, INAOE ext.1303 email: malquori@inaoep.mx

Gustavo Rodríguez Gómez , INAOE ext. 8317 email: grodrig@inaoep.mx

Miguel Martinez Arroyo, INAOE ext 8118, email: mmtz@inaoep.mx

Estudiantes

Juan Pablo Torres Papaqui, INAOE

Fabián Rosales, BUAP & INAOE

Claudia Judith García Calderón, Instituto Tecnológico de Veracruz & INAOE

Miembros extranjeros

Dr. Ofer Lahav, IoA, Cambridge University. +44 (0)1223 337540 . lahav@ast.cam.ac.uk

Dr. Sandro Bressan, Observatorio di Padova. bressan@pd.astro.it

Miembros consultores

Dr. José Luis Zechinelli Martini, UDLA. Tel. (222) 229 26 22. zechinel@mail.udlap.mx