Design and Implementation of Distributed Information Retrieval System Based on VOTable Protocol

Yuyang Hu\textsuperscript{1,a}, Xicheng Qian\textsuperscript{1,b} and JingChang Pan\textsuperscript{1,c,*}

\textsuperscript{1}School of Mechanical, Electrical & Information Engineering, Shandong University (Weihai) 264209, China
\textsuperscript{a}huyuyang_2015@outlook.com, \textsuperscript{b}xichengqian@gmail.com, \textsuperscript{c,*}pjc@sdu.edu.cn

*Corresponding author

Keywords: distributed, sharing, VOTable, retrieval

Abstract: Astronomy has entered a period of massive data with rapid development of the information age. Astronomical research is becoming more and more mature. Nevertheless, there is still a big problem with the sharing massive astronomical data. Because the astronomical data is widely distributed in various observatory stations around the world and the structure, format and language of astronomical data in different regions and even different countries are different, the retrieval and storage of astronomical data is facing great challenge. This paper designs and implements a distributed information retrieval system based on VOTable protocol. In this way, we can improve the sharing of massive astronomical data greatly.

1. Introduction

The explosive growth of astronomical data has brought challenges of the storage, access and retrieval of data, and it is becoming more and more important to solve the distribution problem of astronomical data. So virtual observatory (referred to as "VO") was born\textsuperscript{[1-3]}. VO is a data-intensive online astronomical research environment that integrates seamlessly and transparently with astronomical research resources on a global scale through advanced information technology\textsuperscript{[4]}. In recent years, the development of VO in various countries has become faster and faster. The British VO has established a regional complete virtual astronomy system, AstroGrid, which not only completed the development of the entire virtual terrestrial architecture at all levels, but also considered several scientific applications as its examples\textsuperscript{[5]}. After years of research, according to the relevant standards of the International Virtual Observatory Union (IVOA), China Virtual Observatory uses Globus Toolkits4 (GT4) and other network technologies to eliminate and initially implement its own data access platform (VO-DAS)\textsuperscript{[6-7]}. Based on International Virtual Observatory Union(IVOA) astronomical standard data format VOTable, we established a distributed astronomical information management and retrieval system to provide a transparent variety of heterogeneous data storage, query, maintenance and management system which reduce the difficulty of retrieval operations, improve the efficiency of astronomical information retrieval for the majority of astronomical researchers.

2. System detailed design

2.1 Traditional retrieval method

Fig.1 is the traditional astronomical information retrieval method diagram.
The traditional retrieval method is that each of the astronomical sites retrieves the internal resources, and then returns the data over the network.

### 2.2 Architecture

Based on the structure of the virtual observatory concept system, the architecture of the distributed information retrieval system is designed combining with the practical application requirements and the characteristics. System architecture diagram is shown in Fig.2.

As shown in Fig.2, the system contains three parts: **user**, **node** and **resource**.

1. **User** is responsible for sending the request command to the master node, and mainly retrieving the data.
2. **Node** is divided into master node and slave nodes. The master node is responsible for receiving the commands sent by the user and distributed to slave nodes through the multi-threading mechanism. The slave nodes receive the commands which the master node to analyze and collate and retrieve the data from the resource through the Simple Cone Search (SCS).
3. **Resource** are distributed in each distribution node in the form of data entity files, including star catalog files, astronomical documents, and astronomical materials, for the discovery and processing of each slave node. It is easier to manage that resources are stored in a file rather than a database. The data is returned to slave node in VOTable format, and the slave node returns to master node.
2.3 System implementation

This system implementation method is shown in Fig.3.

![VOTable-based distributed information retrieval method diagram](image)

Figure 3. VOTable-based distributed information retrieval method diagram.

1. Obtaining a summary list of resources. As shown in Fig.3, the system will first obtain a summary list of resources after login, and the summary list of resources is shown in the form of main class - subclass - node. The main class is divided into three types of documents: star catalog, astronomical documents, and spectra, respectively marked with 1, 2, and 3 numbers. Subclass is the file type, such as dat, doc, docx, caj, pdf, etc. Slave nodes are replaced with their own IP. When you click on a node's IP, you can see all the data of this node.

2. Send SCS command. As shown in Fig.3, the master node first sends resource retrieval command to slave nodes.

3. Resource scheduling strategy. In this system, the resource scheduling method is different from the traditional way. The slave nodes will not go to find their own internal resources, but to retrieve the summary list of resources.

4. Returning the search results. Results finally are pooling to the master node and then are presented to the user by the master node.

2.4 Key Technologies

1. **REST architecture.** REST solves the connection problem between the master node and the slave node in the system. In this system, for example, REST uses a URI to represent a resource, such as a star catalog. Then, the presentation layer between the master node and the slave node is the data entity file. The final master node passes the Http's four (Post, Get, Put, Delete) to operate on the slave node resources to achieve "presentation layer state transition."

2. **Callable + Future Multi-threading Technology.** The use of this multi-threaded technology is based on two points. One is to determine whether the task is completed. Second, you can get the implementation results of multi-threaded results, that is, the system's resource search results.

3. **Simple Cone Search (SCS).** SCS is based on three conditions of right ascension, declination, and retrieval radius. Given the coordinates (α, δ) and the retrieval radius (r) on the celestial sphere, we find all the celestial object information with the angular distance of (α, δ) less than r.

4. **VOTable data format.** As shown in the Fig.5, VOTable is composed of Metadata, TableData. Metadata is a metadata that contains two parts, one is the metadata of the form's own information (Parameters, Infos, etc.) and the other is the metadata information (Fields or Groups in...
the figure) for each column of the table data. Fields contains a number of Fields elements, which are
described for each column of data elements. Fields are exactly the same as the number of data
columns. Each Field element contains several attributes and sub elements, describing the name of
the corresponding column (name attribute), ID (ID Attribute), UCD (ucd attribute), data type, data
size and other information, such as a Field element name attribute is DE, then the corresponding
column of the column name DE. Table data is composed of a series of rows and columns, rows and
columns are composed of cells.

```xml
<IDOCID SYSTEM "http://ls-wo.org/html/VOTable.dtd">
  <VOTABLE>
    <DESCRIPTION>
      VO Table
    </DESCRIPTION>
    <RESOURCE>
      <TABLE name="Table1">
        <FIELD id="field1" datatype="double" name="Planet" arraysize="10"/>
        <FIELD id="field2" datatype="double" width="5" name="Diameter"/>
        <GROUP ID="group1" name="PlanetInfo">
          <FIELD ref="field1"/>
          <FIELD ref="field2"/>
        </GROUP>
      </TABLE>
      <DATA>
        <TABLEDATA>
          <TR>
            <TD>45.835330974</TD>
            <TD>7.8798857212</TD>
          </TR>
          <TR>
            <TD>146.8295159796</TD>
            <TD>20.6830716011</TD>
          </TR>
        </TABLEDATA>
      </DATA>
    </TABLE>
  </RESOURCE>
</VOTABLE>
```

Figure 5. VOTable data format.

2.5 Practice

(1) Summary list of resource is shown in Fig.6. The summary list of resource, which is displayed in
form of main class - subclass – node, shows all the resource information about all nodes.

![Table](image)

Figure 6. Summary list of resources.

(2) Node's listening mechanism is shown in Fig.7. When the user click to “Listening” button, it
will show operation status of this site.

![Listening](image)

Figure 7. Node's listening mechanism.

(3) As shown in Fig.8, it is the result of SCS with an angular distance of (45, 90) of less than
0.01 by inputting the right ascension 45, the declination 90 and the search radius 0.01.
3. Summary

We designed and implemented an information retrieval system which is able to build a summary list of resource and then retrieve the summary list of resource. Related to traditional retrieval method, this resource scheduling strategy is more efficient and convenient. The future research work will focus on adding variety of resources and so on.

Acknowledgement

This work was financially supported by the National Natural Science Foundation of China (U1431102)

References


