Yanjie Fu

School of Information Science and Engineering
Graduate University of Chinese Academy of Sciences

Center for Space Science and Applied Research
Chinese Academy of Sciences

Office: +86-010-62586433
Mobile: +86-18611199228
Fax: +86-010-62586433
Email: fu_yanjie@cssar.ac.cn
Homepage: http://sites.google.com/site/pipifuyj/

Biography

Yanjie Fu received his B.E. degree in Computer Science from University of Science and Technology of China. In 2008, he was recommended as an exam-free graduate student, and now is a M.S. candidate in School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences.

Education

Master(2008.09-2011.07)

School of Information Science and Engineering,
Graduate University of Chinese Academy of Sciences (GUCAS).

Bachelor(2004.09-2008.07)

Department of Computer Science,
University of Science and Technology of China (USTC).

Research Interests

Data Mining;

Data Grid, Large Scale Data Sharing Technologies, including their application on satellite data;

Information System, Virtual Observatory.

Honors & Awards

IEEE WCCI Active Learning Competition Winner On Hand Writing Recognition(Rank 1), 2010

Outstanding Graduate Scholarship, University of Science and Technology of China, 2007-2008

Outstanding Graduate Scholarship, University of Science and Technology of China, 2006-2007

Outstanding Student Scholarship, University of Science and Technology of China, 2005-2006

Outstanding Freshman Scholarship, University of Science and Technology of China, 2004-2005

Outstanding Student of USTC Summer Research Project, 2006/07-2006/09

English Certifications

College English Test-Bank4: passed

College English Test-Bank6: passed

GRE: V:340/800 Q:800/800 AW:4/6

Academic Activities

2010.07 The 10th Symposium of Scientific Databases And Information Technology (Guizhou,China)

2010.07 NASA Data Management and Data Interoperability Workshop - Planetary Data System (Shandong,China)

2010.05 AISTAT Active Learning Workshop (Sadinia,Italy)

Presentation: Clustering and Association in Active Learning

2009.11 Chinese Virtual Observatory Workshop (Chongqing,China)

Presentation: The Architecture Of Space Science Virtual Observatory

2009.10 National Physics Conference (Beijing,China)

Publications

Workshop Papers

(1) Yanjie Fu, Ziming Zou, Jizhou Tong, Architecture Research of Space Science Virtual Observatory Based On Autonomic Grid, The 10th Symposium of Scientific Databases And Information Technology (PDF in Chinese)

Projects

Service Oriented Architecture and its applications on Web Service

1. My Donor Demo

   Aims

   With the increasingly growing large scale data of alumni, on one hand, people need to manage detailed records for donors, prospects, volunteers, memberships, alumni and more. On the other hand, to raise more money, it is quite necessary and important to track alumni grants,provide online fundraising approach and propose capital campaigns. Finally and also the most important thing is how to dig out those potential kind-hearted alumni and keep in touch with them. So, that is why we need "my donor".

   Programming Languages and Tools

   PHP+XML+ExtJS+MYSQL+ Statistics Method

   Sponsor

   USTC Initiative Foundation

2. Ways to Give

   It is not always a good way to provide only call-in donation. With the development of e-business, an integrated online donation front end is significant to provide easy ways for alumni to donate. And it is a project supported by USTC Initiative Foundation.

   Sponsor

   USTC Initiative Foundation

3. Lightweight Php Web Framework Source Code

We are always using other Web Framework to create web service or online system. However, we never think about if we can also create a new lightweight Web Framework, even our framework is still not so strong. But, we can understand what is the core mechanism of a web framework. And it is supported by personal interests with my dear friends in Beijing.

It is a Model-View-Controller framework. Our main idea of this framework is as follows:

(1)class "framework" has models+controllers+views;

(2)obtain the action name and method name according to the URL;

(3)require the relative controller PHP file and create an object for this controller class;

(4)require the relative view PHP file and create an object for this view class;

(5)we fire the events according to a single-line order - "beforeInitiationafterInitiationbeforeRenderafterRenderbeforeOutputafterOutput"

(6)we view model as an object: data+actions. model has fields and records. Each record means a row in a data matrix.

Data Mining

1. IEEE WCCI 2010 Competition Program - Data mining (IJCNN) on Hand Writing Recognition Source Code Award Technical Report Workshop

Challenge Protocol

The participants are allotted a budget of "virtual cash" allowing them to "purchase" all the training data labels at the price of 1 ECU (experimental cash unit) per label. They can place queries to the server by providing a list of samples for which they desire to purchase the label. Upon receipt of the labels, their account of virtual cash is debited. The participants are free to choose the number of queries and the number of samples per query. An experiment terminates when all the budget is spent or the challenge deadline is reached. To monitor progress, the participants are asked to provide predictions for all the labels every time they place a query, including the known and unknown labels of the training examples and the labels of the test examples.

When query, an predict and a sample should be done:

Predict - Using the examples with known labels (at first use the seed example, which has a positive label), train a predictor and provide prediction scores for ALL the examples of the dataset (including those used for training). Any sort of numeric prediction score is allowed, larger numerical values indicating higher confidence in positive class membership.

Sample - Choose among the remaining unlabeled examples those for which you want the purchase labels. You may only query examples in the first half of the dataset (training examples). If you query test examples, you will not receive those labels (and not be charged either).

The goal is to purchase as few labels as possible with "virtual cash" while getting as good performance as possible.

Data Set

There are 20000 instances in which every instance have 12000 attributes and 1 unknown label what should be predicted. The possible values for the label are 1 and -1. At first, we know the first instance's label is 1.

Programming Language and Tools

java+python+extweka+libsvm

Result

Award

Winner(Rank 1) on Hand Writing Recognition

2. Space Weather Prediction and Data Analysis for Satellite Orbit

Aims

In the past, physics scientists present their space weather models based on physics theory and use those models to predict space weather and space events. However, the increasingly growing records from satellites leads to a problem that data is over-loaded for scientists. So, how to process large scale satellite data? We should seek help from data mining technologies.

Now, for example, we can divide our data into two kinds: A is data from all kinds of satellite monitors; B are records from historical space events. But A is related to B. And we can view relationship between A and B as C. Consequently, can we dig out the patterns from A+B+C by mathematical approaches? And then translate those patterns into predictors to predict potential space events for specific possibility? Maybe the results are not totally accurate. But they are meaningful.

So this project aims to forecast space event and provide potential satellite-disaster warming for Air Force. The business flow is as follows:

(1)Obtain real-time data from space weather monitoring satellite

(2)Data preprocessing including: error correction, data dimension reduction by sampling frequency, intermediate physical variables computing

(3)Based on historical space weather event records, forecast potential space weather disaster by data mining method like SVM, Classification, Neural Network. (What I mainly focus)

Sponsor

National 973 Plan

3. Finding Topics in A Collection of Documents Source Code Technical Report

Aims

In our daily life, every document has its topic- what this document is mainly talking about. And machine can tell the difference of documents' topics from composition of words. For example, if most words in a document are relative with sports, it is quite possible its topic is sports. So, the question is how to find out the topics in a collection of documents without training dataset. Only by simple clustering? Of course not.

In our work, we use two methods: Kmeans and Shared Nearest Neighbor, and then improve the classic Shared Nearest Neighbor Approach based on Google Page Rank, and compare the clustering accuracy with Kmeans and classic Shared Nearest Neighbor approach.

Programming Languages and Tools

java

Result

Database and Large Scale Data Sharing Technology

1. Database and Large Scale Data Sharing Technology Demo

Sponsor

Chinese National Eleventh Five-Year Plan

Programming Language and Tools

PHP+ExtJS+Google API+MYSQL+JAVA+JSP

Resource Layer

Space science data is quite different from other common data. Frequently, space science data is stored in CDF format or Plain Text format or FITS format. At the same time, space science data is always located in different stations, observatories and institutes. So, in some way, it is a distributed file system. But to search, locate and retrieve needed data as quickly as possible, we need a metadata model or a metadata standard. We can divide metadata into two kinds by its formats. One is metadata in file; another is metadata in database. Another problem is how to design metadata model including category, instrument, equipment, space region and so on. To make the metadata model properly suite the usage of space science community, we need to take a lot of factors into consideration.

So, a dataset should contain data, standard metadata and core metadata. Here, data support physics experiments and analysis; standard metadata means metadata in file format which describe the detailed information of data, and it supports interoperability; core metadata means metadata in relative database which only contains important core attributes, and it supports data discovery. Data provider is a larger concepts including many datasets. Resource Layer provide data access and storage.

Service Layer

Resource Register and Delivery

Resource Audit

Resource Discovery

Metadata Harvest and Synchronization

Data and Service redundancy

Transparent Switching

Single-Sign On

Work Flow For Satellite Data Preprocessing

Service Layer

Space Weather Model Computing

Google Earth Visualization Service

Atmosphere Density and Temperature Models

Space Weather Event Association Data Discovery

Last updated: July 19, 2010