基于频域变换的森林生态站观测指标聚类分析

计算机软件与理论

3130295 张兆玉

导师 陈志泊
导师 王建新

目录

一、选题的理论和实践意义

- 研究背景

中国的生态系统研究网络（CERN）于1988年开始组建成立。目的是为了监测中国生态环境变化，综合研究中国资源和生态环境方面的重大问题，发展资源科学、环境科学和生态学。森林作为人类文化的摇篮和绿色宝库，是重要的可再生资源，无疑使得森林生态站的存在举足轻重。

通过几十年对研究方向的专注和探索，我国的森林生态站的发展已经取得长足的发展，并积累了大量丰富的观测数据。发展至今，我们面临的问题是该如何利用这些宝贵的数据，来发现其潜在的价值。

一、选题的理论和实践意义

- 数据挖掘学科的发展

数据挖掘（Knowledge Discovery in Data 简称：KDD）技术的出现为我们从浩如烟海的数据中挖掘黄金提供了可能。利用数据挖掘的一般过程：数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估，到最终发现有用的知识（即模式），我们可以对生态站的数据进行科学有效的分析。

数据挖掘的功能主要包括：1. 类/概念描述：特征化与区分；2. 挖掘频繁模式、关联与相关性；3. 用于预测分析的分类与回归；4. 聚类分析；5. 离群点监测。

一、选题的理论和实践意义

面临问题

通过对现有的森林生态站观测指标数据的预处理和分析，发现如下问题：

- 缺乏一个统一的平台来对观测的数据进行管理，导致数据的价值没有被最大化利用。数据的共享严重不足。

- 由于观测仪器老化、断电、系统故障等原因导致数据缺失，是否可以根据数据之间的相关性，进行数据补全工作。

- 观测指标的观测值是基于时序的，各观测指标间是否存在某种时间周期上的相关性。

一、选题的理论和实践意义

- 实践意义

通过将基于离散时间傅里叶变换和QENNI算法引入到森林生态指标数据的分析中，不仅可以揭示各个指标间内在的联系，为下一步数据的分析奠定基础。同时，通过QENNI数据补全算法能够保证观测指标数据的正确性和准确性，也为聚类算法建立模型提供了精准的数据基础。

目录

二、数据挖掘技术现状与趋势

- ## 数据挖掘的现状

在第11届国际联合人工智能学术会议上KDD被首次提出。截止到现在，数据挖掘领域已基本成熟。因此，KDD国际会议研讨会的研究重点已经从方法过度到应用。

在国外，商务智能（BI）领域，利用数据挖掘提供商务运作的历史、现状和预测视图，包括业绩管理、标杆管理和预测分析。特别是在电子商务领域的用户个性化推荐；在制造业中，半导体的生产和测试中都产生大量的数据，就必须对这些数据进行分析，找出存在的问题，提高质量；在生物领域，采用数据挖掘的手段对DNA进行分析；在银行和保险行业中通过离群点检测对发生的异常行为进行检测。另外，机器学习、数据库、数据仓库、人工智能、信息检索等领域的国际学术期刊业都先后开辟了数据挖掘专题或专刊。当前，国外的数据挖掘主要集中在对知识发现方法的研究，通过将不同学科的新方法的融合来显著增强数据挖掘的能力。

而在应用方面通过商业数据挖掘软件的不断产生和完善，针对问题的领域建立一个系统的整体方面是现在发展的方向。

二、数据挖掘技术现状与趋势

- 数据挖掘的现状

在中国，数据挖掘的起步要晚于国外。但是，学习数据挖掘基础理论和应用研究的人越来越多，也使得中国的数据挖掘领域正在飞速地前进。目前，在我国清华大学、中科院计算机研究所、人民大学等在内的很多单位都已经开设数据挖掘相关的学习课程。

数据挖掘的应用，包括天猫、京东等电商平台的对用户数据的挖掘，百度、搜狗等搜索公司也对用户数据和日志数据进行挖掘，从中发现很多有价值的模式。在未来，数据挖掘肯定会渗透到越来越多的领域，跟行业特点紧密相关挖掘方法肯定会越来越多地被运用。

二、森林生态站数据挖掘系统的研究现状

数据挖掘的现状

国内外森林生态站中数据挖掘技术的应用往往比较单一。分类模式、关联规则等挖掘方法仅仅单一的使用在生态站数据中。

在未来的研究中多种方法相结合系统地运用到森林生态站数据挖掘中，可以发现很多原本隐藏的模式，用来揭示数据内在的规律，必定可以用来解决很多实际中遇到的问题。

森林生态站数据管理面临的问题

- 森林生态站观测的数据比较分散。
- 数据量巨大，传统的文件式存储已难以满足现实的需求。
- 依托计算机的信息化时代早已来临，很多生态站基本上还是半手工甚至是纯手工的状态来整理管理这些数据。。
- 在当前的工作状态环境下，全国各地的生态站之间并没有建立起统一的数据管理与共享平台。各森林生态站仍就是一个个信息孤岛。
- 随着数据量的增加，新的应用情景下需要新的数据挖掘算法。

二、森林生态站数据挖掘系统的研究现状

数据挖掘的趋势

- 统一的多站合作的数据共享管理平台的建立。
- 基于分布式存储的大数据存储方案运用到森林生态站数据管理平台。
- 多种算法相结合的数据挖掘方法将被更多的应用到森林生态站数据的挖掘工作中。
- 专门用于知识发现的数据挖掘形式化的描述语言DMQL将走向形式化和标准化。
- 可视化数据挖掘过程，能够使得数据的挖掘结果很容易的被使用者理解。
- 通过数据挖掘所发现的各种模式将会成为系统服务的一部分，面向普通用户公布，通过这个途径可以将数据转化为市场价值。
- 实时数据挖掘。
- 预测性数据挖掘。

目录

三、研究内容及目标

- **研究内容及目标**
  - **基于皮尔逊积矩相关系数**（Pearson product-moment correlation coefficient简称：PPMCC）的观测指标相关性研究。通过PPMCC算法，能够得到观测指标之间的相关性，为数据填充算法选取样本数据的理论依据。
  - **加权的基于象限最近邻填充算法**（Weighted Quadrant Encapsidated Nearest Neighbor based Imputation 简称：WQENNI）的数据补全研究。
  - **基于离散时间傅里叶变换的聚类算法**（Discrete-time Fourier Transform Clustering Algorithm 简称：DTFTCA）森林生态站观测指标聚类分析。
  - **森林生态站数据管理平台**（Forest Ecology Station Data Management Platform简称：FESDMP）

目录

四、研究方法、技术路线、实验方案

- 数据样本的选取

➢ 选取不同年份的相同时间段的指标的离散观测值。比如缺失2014年7月1日12:00的观测值，则选取2004年~2014年的7月1日观察的观测值作为相关数据。

➢ 依据皮尔逊积矩相关系数相关性分析，选取n个具有相关性的观测指标，获取该n个指标2014年7月1日的所有观测数据，作为QENNI算法的输入数据。

- 皮尔逊积矩相关系数

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

四、研究方法、技术路线、实验方案

- 加权的基于象限近邻填充算法（WQENNI）

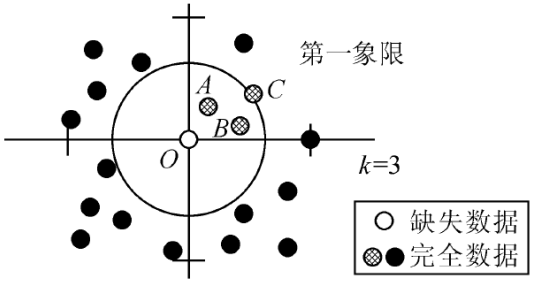


图1　kNN 算法最近邻点的选择

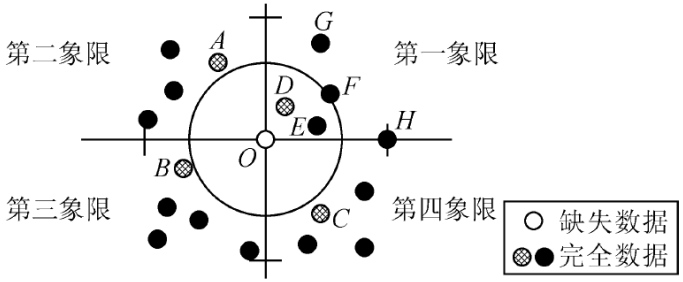


图2　QENNI 算法最近邻点的选择

- 算法改进
  - 距离平方的倒数作权值
  - 方向距离综合加权，降低数据过多的分布在某几个象限而仅仅选取一个值对数据的准确性的影响。

四、研究方法、技术路线、实验方案

- 离散时间的傅里叶变换

➢ 傅里叶变换是将信号在时域（或空域）和频域之间变换时使用。对信号进行时域分析时，有时一些信号的时域参数相同，但并不能说明信号就完全相同。因为信号不仅随时间变化，还与频率、相位等信息有关，这就需要进一步分析信号的频率结构，并在频率域中对信号进行描述。该定理表明任何连续测量的时序或信号，都可以表示为不同频率的正弦波信号的无限叠加。

➢ 森林生态站指标的观测数据都是基于时序的数据，很难直接在时域上通过距离公式发现其之间潜在的相似周期性变化。

- 离散时间傅里叶变换（**DTFT，Discrete-time Fourier Transform**）

四、研究方法、技术路线、实验方案

离散时间傅里叶变换 离散时间傅里叶变换（DTFT，Discrete-time Fourier Transform）以离散时间 nT (其中n ∈ Z，T 为采样间隔) 作为变量函数（离散时间信号）$f(nT)$变换到连续的频域，即产生这个理算时间信息的连续频谱 $F(e^{i\omega})$，值得注意的是这一频谱是周期的。

记连续时间信号 $f(t)$的采样为：

$$f_{sp}(t) = \sum_{n=-\infty}^{\infty} f(t)\delta(t - nT)$$

公式七

其傅里叶变换为：

$$\mathfrak{F}\{f_{sp}(t)\} = \int_{-\infty}^{\infty} f_{sp}(t)e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(t)\delta(t-nT)e^{-i\omega t}\,dt = \sum_{n=-\infty}^{\infty} f(nT)\,e^{-in\omega T}$$

四、研究方法、技术路线、实验方案

公式八

这就是采样序列 $f(nT)$的 DTFT：

$$F_{DTFT}(e^{i\omega T}) = \sum_{n=-\infty}^{\infty} f(nT)\, e^{-in\omega T}$$

公式九

为方便起见，通常将采样间隔 T 归一化，即为 $f(nT)$的离散时间傅里叶变换：

$$F_{DTFT}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f(n)\, e^{-in\omega}$$

公式十

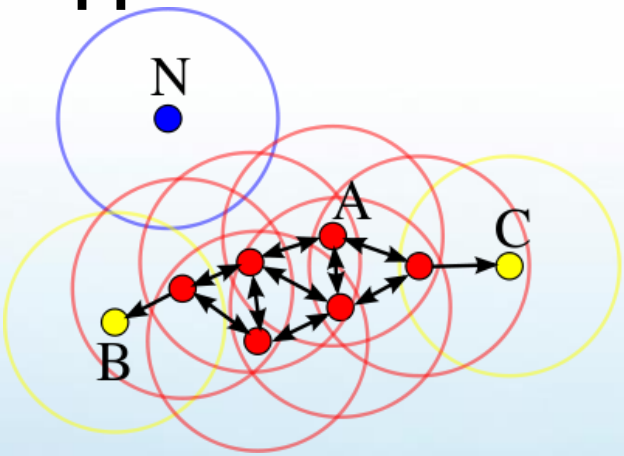
其反变换为：

$$f(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_{DTFT}(e^{i\omega})\, e^{in\omega}\, d\omega$$

公式十一

四、研究方法、技术路线、实验方案

- 基于密度聚类的算法（**Density-based Spatial clustering of applications with noise** 简称 **DBSCAN**）



- 算法核心思想：

从某个选定的核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

四、数据挖掘技术现状与趋势

- 基于离散时间傅里叶变换的聚类算法（**Discrete-time Fourier Transform Clustering Algorithm** 简称**DTFTCA**）

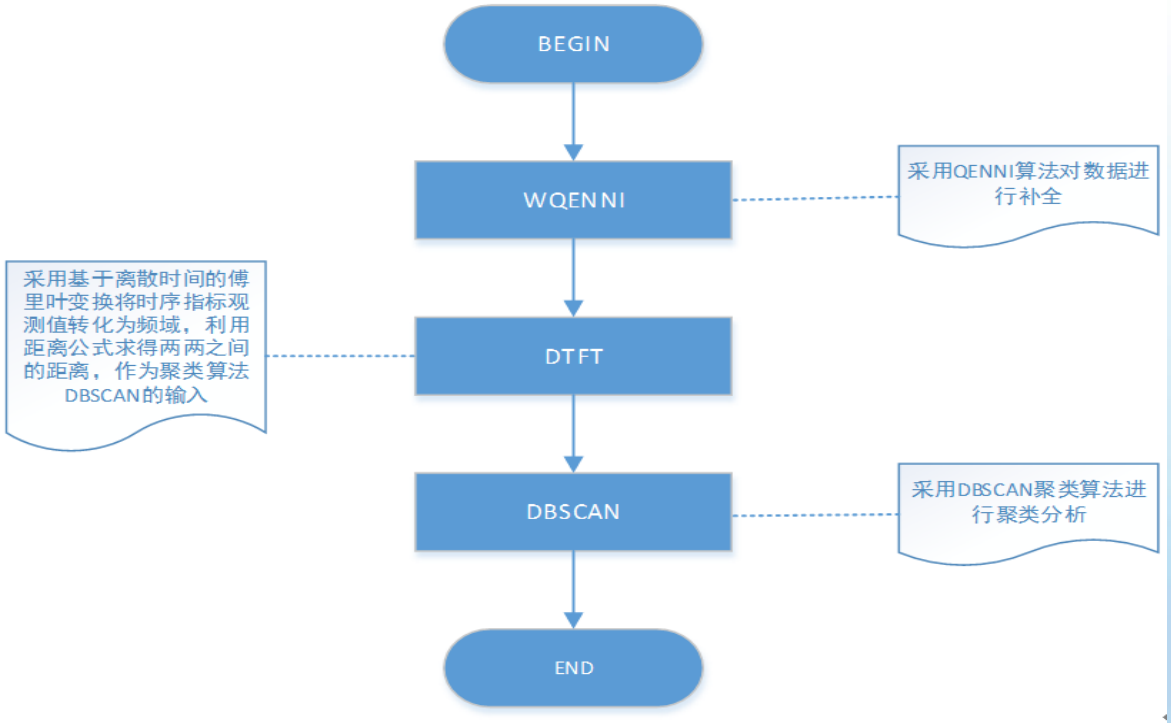


图 5 聚类分析算法整体流程图

四、数据挖掘技术现状与趋势

- 聚类算法结果分析图

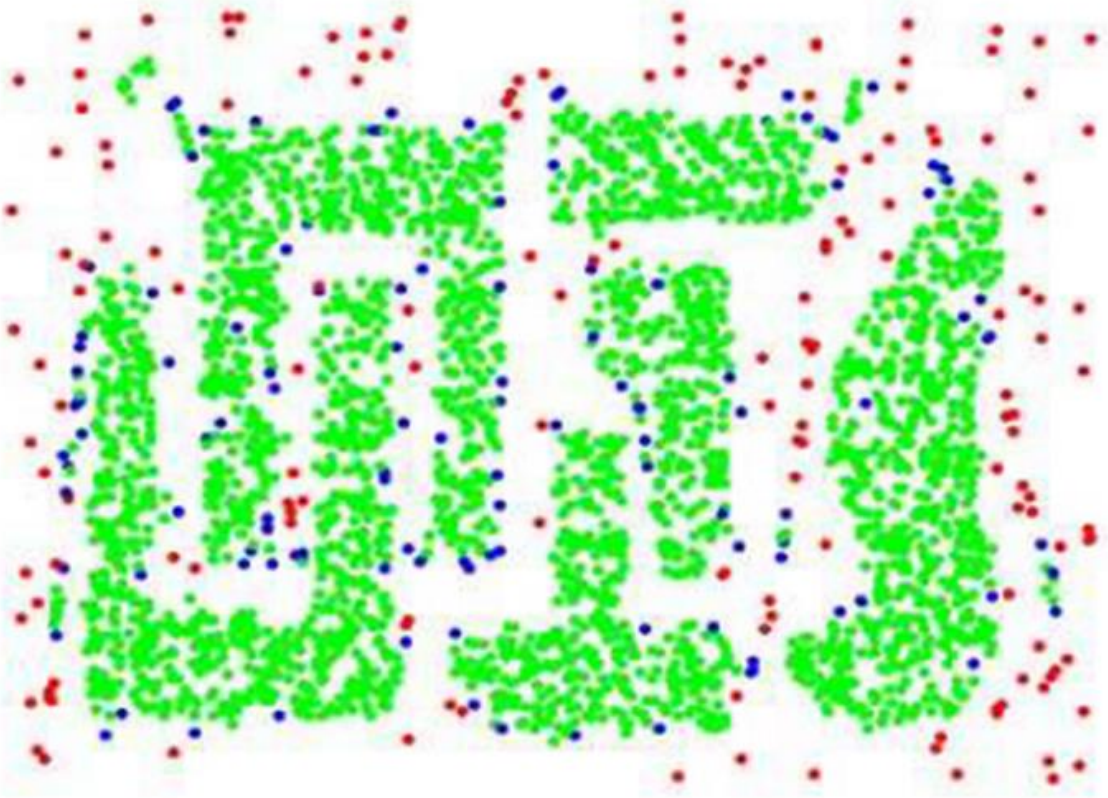


图 6 聚类分析结果图

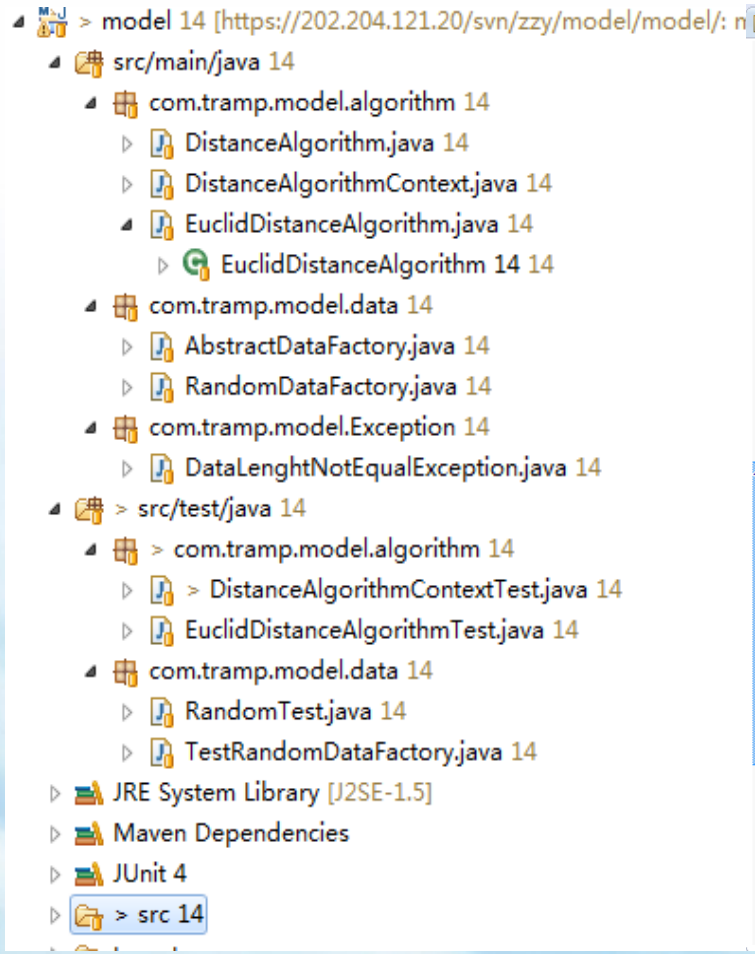
目录

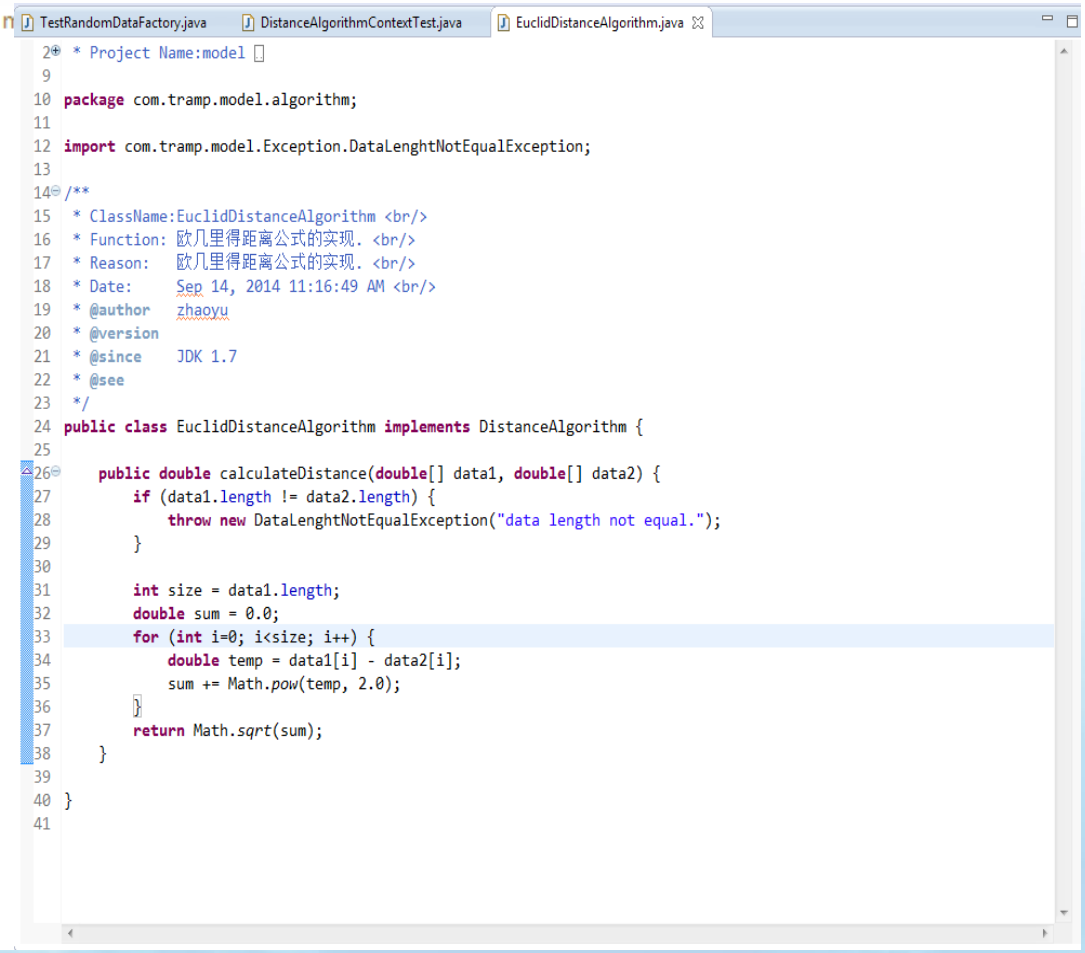
五、可行性分析

可行性分析

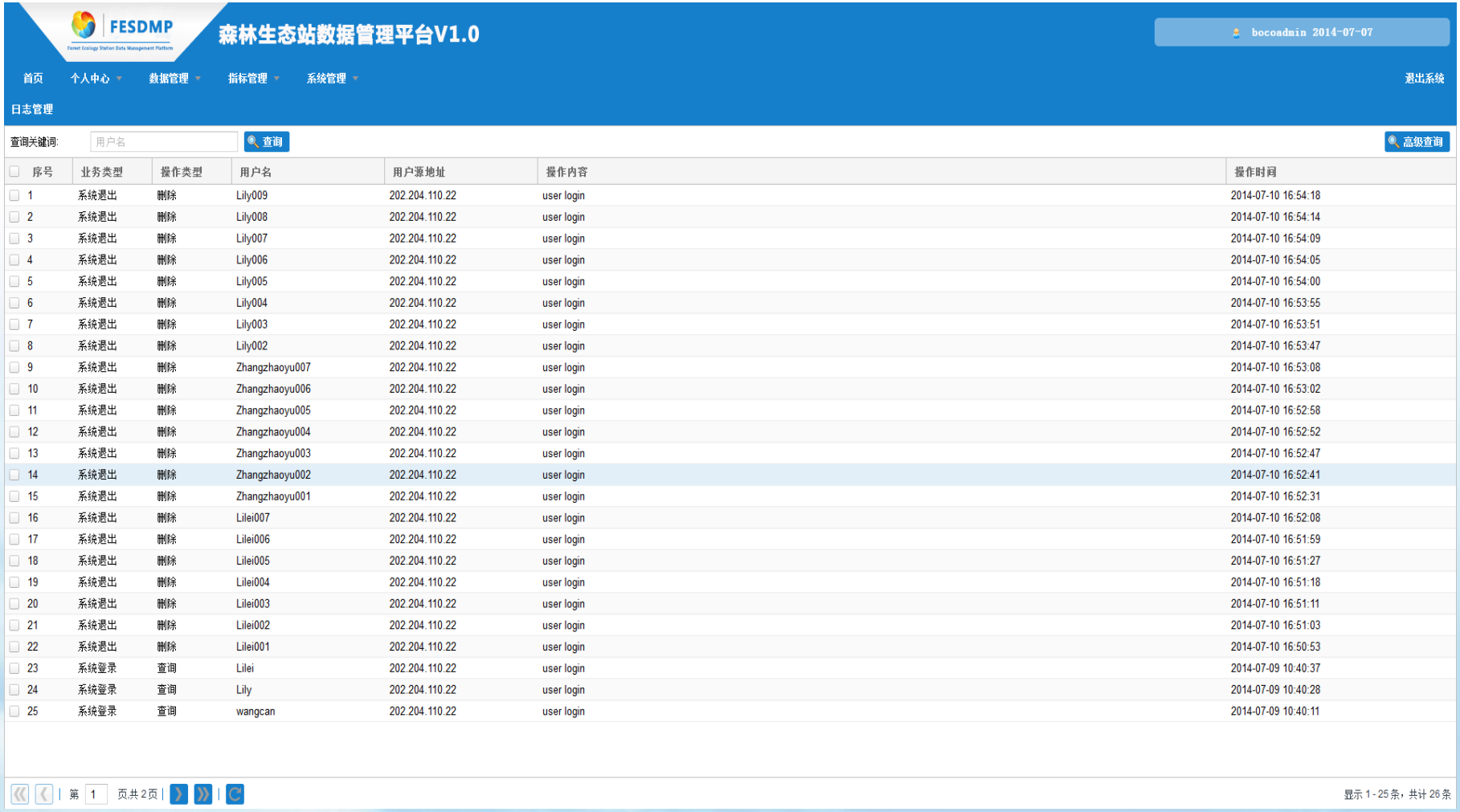
- 数据挖掘是一个动态的、强势快速扩展的领域。
- 在理论方面，通过阅读大量相关文献，已经了解到数据挖掘应用的热门研究策略，并在现有的数据填充策略基础上进行了优化研究制定了新的研究方法，并针对森林生态站基于时序的数据引入傅里叶变换聚类方法。
- 技术方面，算法程序采用**Java**编程语言。本人有近四年的**Java**编程开发经验，相信在算法的实现上没有阻碍。
- 项目建设上，本人有相对丰富的项目开发、管理经验。从事过北林数字标本馆、成人教育学院学籍管理系统等众多项目的开发。相信在工程实施上能在保证质量的前提下，按时完成。

五、可行性分析


> model 14 [https://202.204.121.20/svn/zzy/model/model/: n
src/main/java 14
com.tramp.model.algorithm 14
DistanceAlgorithm.java 14
DistanceAlgorithmContext.java 14
EuclidDistanceAlgorithm.java 14
EuclidDistanceAlgorithm 14 14
com.tramp.model.data 14
AbstractDataFactory.java 14
RandomDataFactory.java 14
com.tramp.model.Exception 14
DataLenghtNotEqualException.java 14
> src/test/java 14
> com.tramp.model.algorithm 14
> DistanceAlgorithmContextTest.java 14
EuclidDistanceAlgorithmTest.java 14
com.tramp.model.data 14
RandomTest.java 14
TestRandomDataFactory.java 14
JRE System Library [J2SE-1.5]
Maven Dependencies
JUnit 4
> src 14


TestRandomDataFactory.java
DistanceAlgorithmContextTest.java
EuclidDistanceAlgorithm.java
* Project Name:model
package com.tramp.model.algorithm;
import com.tramp.model.Exception.DataLenghtNotEqualException;
/**
* ClassName:EuclidDistanceAlgorithm
* Function: 欧几里得距离公式的实现.
* Reason: 欧几里得距离公式的实现.
* Date: Sep 14, 2014 11:16:49 AM
* @author zhaoyu
* @version
* @since JDK 1.7
* @see
*/
public class EuclidDistanceAlgorithm implements DistanceAlgorithm {
public double calculateDistance(double[] data1, double[] data2) {
if (data1.length != data2.length) {
throw new DataLenghtNotEqualException("data length not equal.");
}
int size = data1.length;
double sum = 0.0;
for (int i=0; i<size; i++) {
double temp = data1[i] - data2[i];
sum += Math.pow(temp, 2.0);
}
return Math.sqrt(sum);
}
}

五、可行性分析



序号	业务类型	操作类型	用户名	用户源地址	操作内容	操作时间
1	系统退出	删除	Lily009	202.204.110.22	user login	2014-07-10 16:54:18
2	系统退出	删除	Lily008	202.204.110.22	user login	2014-07-10 16:54:14
3	系统退出	删除	Lily007	202.204.110.22	user login	2014-07-10 16:54:09
4	系统退出	删除	Lily006	202.204.110.22	user login	2014-07-10 16:54:05
5	系统退出	删除	Lily005	202.204.110.22	user login	2014-07-10 16:54:00
6	系统退出	删除	Lily004	202.204.110.22	user login	2014-07-10 16:53:55
7	系统退出	删除	Lily003	202.204.110.22	user login	2014-07-10 16:53:51
8	系统退出	删除	Lily002	202.204.110.22	user login	2014-07-10 16:53:47
9	系统退出	删除	Zhangzhaoyu007	202.204.110.22	user login	2014-07-10 16:53:08
10	系统退出	删除	Zhangzhaoyu006	202.204.110.22	user login	2014-07-10 16:53:02
11	系统退出	删除	Zhangzhaoyu005	202.204.110.22	user login	2014-07-10 16:52:58
12	系统退出	删除	Zhangzhaoyu004	202.204.110.22	user login	2014-07-10 16:52:52
13	系统退出	删除	Zhangzhaoyu003	202.204.110.22	user login	2014-07-10 16:52:47
14	系统退出	删除	Zhangzhaoyu002	202.204.110.22	user login	2014-07-10 16:52:41
15	系统退出	删除	Zhangzhaoyu001	202.204.110.22	user login	2014-07-10 16:52:31
16	系统退出	删除	Lilei007	202.204.110.22	user login	2014-07-10 16:52:08
17	系统退出	删除	Lilei006	202.204.110.22	user login	2014-07-10 16:51:59
18	系统退出	删除	Lilei005	202.204.110.22	user login	2014-07-10 16:51:27
19	系统退出	删除	Lilei004	202.204.110.22	user login	2014-07-10 16:51:18
20	系统退出	删除	Lilei003	202.204.110.22	user login	2014-07-10 16:51:11
21	系统退出	删除	Lilei002	202.204.110.22	user login	2014-07-10 16:51:03
22	系统退出	删除	Lilei001	202.204.110.22	user login	2014-07-10 16:50:53
23	系统登录	查询	Lilei	202.204.110.22	user login	2014-07-09 10:40:37
24	系统登录	查询	Lily	202.204.110.22	user login	2014-07-09 10:40:28
25	系统登录	查询	wangcan	202.204.110.22	user login	2014-07-09 10:40:11

目录

六、研究计划及预期成果

- 研究计划

时间节点	研究内容
2014.03 - 2014.06	阅读文献，查阅资料，选题
2014.07 – 2014.09	算法设计，编写程序，，撰写文献综述及开题报告
2014.10 - 2015.08	优化程序代码，撰写小论文及，完成毕业论文初稿
2015.09 - 2016.04	修改毕业论文，定稿，准备答辩

- 预期成果

优化并实现基于象限近邻填充算法在多维空间上的数据补全
优化并实现基于傅里叶变换密度聚类算法在森林生态站观测数据上的使用
发表EI或核心期刊论文一篇

Thank You !