

**本科毕业设计（论文）**

****

**题目：在线社会学习系统的大数据挖掘**

**姓 名： 褚湘鸣**

**学 院：信息与通信工程学院**

**专 业： 信息工程**

**班 级： 2012211126**

**学 号： 2012210700**

**班内序号： 12**

**指导教师： 纪晓东**

**2016 年 6 月**

**在线社会学习系统的大数据挖掘**

**摘 要**

在线社会学习系统的大数据挖掘研究如何向用户提供与其访问的内容密切相关的信息，从而提高在线社会学习系统内部推送广告的效率。建立一个以用户和关键词为节点的图论模型，通过区分该模型上节点之间关系的紧密程度来把该图分为若干群体。在获得用户所访问的内容后，向用户推送该群体内与其他节点的关联和互动最为频繁的节点，既能保证该信息与用户所需内容有紧密的关联，又能提升广告效率。

本文介绍了为达到该目标所做的工作。首先通过分析在线社会学习系统的论坛文本数据提取关键词。然后通过关联分析建立用户和关键词的关联。建立图论模型时，将这些关联作为图中节点和边的权值。最后通过连通图上的K均值算法进行聚类分析，得到图中的簇，即我们所求的群体。群体中每个都有用来表示和群体内其他节点关联频繁程度的权值，需要时将权值最大的节点推出，即解决了该问题。

该方案类似基于关联规则的个性化推荐系统[12]，其优势在于利用图论模型解决了文本数据较为抽象的问题，以及利用聚类分析将节点分为簇，分析大量数据效率较高。

关键词 在线学习 文本挖掘 关联分析 聚类分析 图论

**Big Data Mining On Social Learning Systems**

**ABSTRACT**

Big data mining on social learning systems researched how to provide a user the most mentioned information by the pages or courses the users was visiting, which enhanced the effectiveness of the internal ads on the system. This article introduced a solution, which was to construct a graphic model with nodes made from users and key words, thus split them into several clusters by the closeness between the nodes. Once acquired the information the user was visiting, the system would push the most frequently related nodes within the cluster, which guaranteed the association between the visited information and the pushed information, as well as the effectiveness of the ads.

This article also showed the way to accomplish the target and the work on it. Firstly, extract the key words by analysis on forum text data. Secondly, set connections between users and key words by association analysis. Thirdly, construct the graphic model by nodes of users and key words and edges of these connections. Last but not least, split the nodes by cluster analysis by K-means algorithm. Every node had a weight represented the frequency of its connection to other nodes in its cluster. While needed, the nodes with maximal weights would be pushed out.

This was considered as a kind of personalize recommendation system based on association rules, of which the advantages lied in the graphical model which handled the abstraction of text data and was more efficient since the data had been split into clusters.

**KEY WORDS** online courses text mining association analysis cluster analysis graph theory

**目 录**

[第一章 绪论 4](#_Toc451304033)

[1.1 课题简介 4](#_Toc451304034)

[1.1.1 课题背景 4](#_Toc451304035)

[1.1.2 研究内容 5](#_Toc451304036)

[1.2 论文结构 5](#_Toc451304037)

[第二章 在线社会学习系统数据分析 6](#_Toc451304038)

[2.1 SLN系统结构 6](#_Toc451304039)

[2.2 SLN数据分析 7](#_Toc451304040)

[2.2.1 点击流数据分析 7](#_Toc451304041)

[2.2.2 论坛文本数据分析 8](#_Toc451304042)

[2.2.3 论坛数据优势 9](#_Toc451304043)

[2.3 SLN图论模型 9](#_Toc451304044)

[2.3.1 图论模型的节点和边 9](#_Toc451304045)

[2.3.2 图论模型算法 10](#_Toc451304046)

[第三章 数据挖掘技术 13](#_Toc451304047)

[3.1 数据挖掘技术简介 13](#_Toc451304048)

[3.2 数据收集:网络爬虫技术 13](#_Toc451304049)

[3.3 数据预处理 14](#_Toc451304050)

[3.2.1 数据清洗过程 15](#_Toc451304051)

[3.2.2 数据提纯过程 15](#_Toc451304052)

[3.2.3 文本分析程序 16](#_Toc451304053)

[3.4 关联分析 16](#_Toc451304054)

[3.5 聚类分析 19](#_Toc451304055)

[3.5.1 聚类分析简介 19](#_Toc451304056)

[3.5.2 K均值算法 20](#_Toc451304057)

[第四章 数据挖掘成果 22](#_Toc451304058)

[4.1 获取数据 22](#_Toc451304059)

[4.2 预处理 23](#_Toc451304060)

[4.3 分词及提取关键词 24](#_Toc451304061)

[4.4 关联分析 26](#_Toc451304062)

[4.5 建立图论模型 26](#_Toc451304063)

[4.6 聚类分析 26](#_Toc451304064)

[4.7 挖掘结果 26](#_Toc451304065)

[第五章 总结 27](#_Toc451304066)

[参考文献 27](#_Toc451304067)

[致 谢 28](#_Toc451304068)

# 第一章 绪论

## 1.1 课题简介

### 1.1.1 课题背景

在线社会学习系统（Social Learning Network，简称SLN）是一种综合了学习者、教学者和学习模式的社交网络。

SLN包含多种应用场景，一般来说是一部分人通过在SLN上的互动向另一部分人学习。在SLN上，用户所扮演的角色和学习的模式是由他们所经历的特定的学习过程来决定的。在这些学习过程中，用户之间将会反复地互相传播知识和信息，教学者和学习者之间会互换角色。在某一个特定的时刻，一个用户传播或给出某种他精通的知识时，他就是教学者；一个用户接受某种他不太熟悉的知识时，他就是学习者。此外，合作学习是SLN的一种重要形式，用户通过合作学习各显其能，一起解决目标问题[2]。这些特点使用户之间的交往模式具有显著的社交网络特征。与此同时，SLN具有很强的线下沉淀，许多网络上的用户也是生活中的朋友。

近年来，SLN的发展十分迅猛，经过相当一段时间的演化形成了一种主要的形式：大规模开放在线课程(Massive Open Online Courses)，简称MOOC。

MOOC是SLN的平台化应用，突出规模化、开放性、不受时空限制等几大特点。表1-1列出了一部分国内外具有较大影响力的MOOC平台。

|  |  |  |
| --- | --- | --- |
| 名称 | 地区 | 说明 |
| Coursera | 美国 | 由两名斯坦福大学计算机教授创办，三巨头之一 |
| edX | 美国 | 由麻省理工和哈佛大学联合创创办，三巨头之一 |
| Udacity | 美国 | 翻转课堂模式，通过认证考试和推荐职位盈利，三巨头之一 |
| 学堂在线 | 中国 | 由清华大学创办 |
| Ewant | 中国 | 由上海交大，西安交大，西南交大，北京交大，台湾国立交大共同组建 |
| Open2Study | 澳洲 | 澳洲最大的MOOC平台 |

表 1-1 部分MOOC平台

这些应用将大量的高质量内容开放给所有用户。人群接受普通高等教育。除此之外，在一些与大学或其他有资质的教育机构合作的MOOC平台上，可以申请相关课程的结业证书，为学校、企业提供了的远程培训服务。

除了平台化的MOOC外，还有一种SLN的重要形式——翻转课堂（flipped classroom），简称FLIP。在FLIP下，宝贵的课堂时间给予学生来专注于个性化、合作性、灵活性较强的项目，课后的时间内由学生自主规划和完成标准化的知识学习过程。如看教学视频、阅读功能增强的电子书、听录音、在网络上与其他同学讨论等。

SLN是一种混合教育方式的强大载体，结合了大量的探究性学习方法，让学生的学习更加主动、灵活，参与度更强。

### 1.1.2 研究内容

随着SLN的发展日新月异，来自不同国家和地区、不同年龄段、不同教育背景的用户源源不断地加入到SLN中，使SLN的数据量呈爆炸式增长。有些热门课程甚至有来自全球的数以十万计的用户在同时观看。

对于一个用户来说，进入SLN后的首先会查找自己感兴趣的课程或其他用户。在达到这个目的之前会在系统上根据自己的兴趣爱好浏览。在这个过程中，系统根据用户浏览的内容进行推荐，帮助用户找到自己想要的内容。

为达到这个目的，应设计一个基于内容的推荐系统。在这个过程中所研究的内容和所完成的工作有：

1. 研究SLN的起源、发展和现状，建立推荐系统的使用场景和用户需求。
2. 研究商务智能推荐系统、网络爬虫技术、文本分析技术、数据挖掘技术以及图论。
3. 研究中文SLN典型案例，分析用户产生内容的组成和结构，确定以论坛为主要数据来源，以大数据挖掘技术中的关联分析和聚类分析为主要分析手段。
4. 建立图论模型，并确定节点和边的组成和权值，以及设计相应的算法。
5. 在目标平台上通过网络爬虫技术获取足量的数据，并对其进行格式化、数据清洗、数据提纯等预处理。
6. 通过文本分析技术进行文本分词和提取关键词。
7. 通过关联分析技术，建立图论模型。
8. 通过聚类分析技术，将用户和信息合理分为若干簇，簇内用户和信息之间联系频繁，簇间用户和信息之间联系稀少。

## 1.2 论文结构

本文各章节整体结构如下：

第一章简要介绍课题和论文的整体情况，其中1.1.2小节列出了所完成的工作。

第二章简要介绍在线社会学习系统的结构并对其建立图论模型。其中2.1节分析SLN的系统结构和用户数据组成， 2.2节重点分别分析用户数据中的重要组成部分：点击流数据和论坛数据，并阐述了选择挖掘论坛数据的原因。2.3节详细介绍图论模型的建立过程。

第三章介绍数据挖掘技术在SLN上的应用。其中3.1节介绍数据收集过程和网络爬虫技术；3.2节介绍数据预处理；3.3节和3.4节分别介绍数据挖掘技术中的两个核心技术：关联分析和聚类分析，并介绍了支持度-置信度框架和K均值算法在图论上的一种应用。

第四章详细介绍在SLN上的数据挖掘所做的具体工作，包括从网站上获取数据、数据预处理、文本分词、提取关键词、关联分析、建模、聚类分析的结果。

# 第二章 在线社会学习系统数据分析

## 2.1 SLN系统结构

图 2-1 SLN功能选项卡



图2-1是学堂在线大学先修课中的“概率论与数理统计（2016春）”课程的功能选项卡。该系统中各模块和功能是：

1. 课件：教学视频
2. 课程信息：课程简介、学习进度、公告等信息
3. 讨论区：用于答疑、资料报错、交流信息等
4. Wiki：用于展示该学科的详细信息
5. 进度：按日期显示用户的学习进度
6. 课程讲义：课程文本资源
7. 导学：教学大纲、章节简介和引导性学习
8. 考试：由教学者或管理者组织的考试，部分课程颁发证书

用户在SLN上最主要的行为就是观看课件和在讨论区中发言，所以“课件”和“讨论区”是绝大部分SLN都包含的功能。SLN的大数据挖掘旨在通过用户数据提升用户使用体验和提升销售效率，其重点工作就是分析用户在观看课件和浏览网页时产生的点击流数据以及分析用户在讨论区中写下的文本数据。

在2.2节中将详细分析点击流数据和坛文本数据，并比较其优劣，说明为何选择挖掘论坛文本数据。

## 2.2 SLN数据分析

### 2.2.1 点击流数据分析

点击流(Clickstream)是用来记录用户在浏览网页或使用其他软件时在屏幕上的点击情况的数据。点击流分析是一种非常实用的用户行为分析方法[3]，广泛应用于软件测试、市场调查以及分析员工生产效率等项目。

点击流数据来源于网站的服务器产生的日志文件，这些日志文件中包括每个访问者产生的所有HTTP请求以及执行情况。这些日志的用途是为网站管理员提供管理数据，通过这些点击流数据，我们可以直观地看到用户在网站上正在干什么[7]。

当一个用户在浏览网站时，他所点击的链接、打开的窗口、使用的功能都可以为我们所捕获、存储到点击流数据仓库中，用于分析网站用户的行为和偏好[4]，优化网站内容和推送交叉销售内容（cross-sell，指根据用户的兴趣推荐相关商品）和提升销售内容（up-sell，指根据用户曾经购买的项目推送更高价格的商品）。点击流数据也可以用来提升用户对该网站和其公司的满意度，提升网站的服务效率[6]。

例如，在SLN上的某类课程中，假设用户在第分钟退出，满足的分布由式（2-1）给出：

 式（2-1）

假设按照教学规律，该课程的视频最少需要5分钟讲清知识点，且想要至少70%的学生能够1次看完。由点击流数据挖掘构建的分布函数得到，当时，，说明如果视频控制在5~10分钟的长度内就能基本满足要求。

点击流数据是使网站不断改进的重要依据，许多类似问题都能通过分析点击流解决。分析点击流的步骤为：

1. 从服务器日志文件中找出用户操作；
2. 还原用户行为构建流数据；
3. 分析导致用户行为的原因；
4. 通过数据挖掘技术找出这些行为和原因的聚类、关联，并检测异常等；
5. 数据存储到点击流数据仓库中。

点击流数据仓库的基础是元模式和多维数据库设计，同时也包括对用户行为的复杂分析。一般来说，会有两种表来记录点击流数据。一张表来记录所有的“点击数据”，就像一条拓扑链中的“点”，另一张表用来把同一个用户的这些“点”数据串联起来，一般具有用户标识、访问来源（入口）、进入时间、离开时间、访问时的点击次数、进入页面、离开页面等字段[7]。

当网站的日志文件转化为这些直观的点击流数据时，针对某一条数据，我们其实已经可以较为明显地分析出用户的操作历史，对于获得常见的网站分析指标来说已经足够了。如果我们的点击流数据仓库中保有大量的数据，利用数据挖掘技术可以得到许多隐藏在其中的关键信息，从而帮助开发设计人员改善用户体验。

然而，检测点击流数据需要从海量的日志文件中查找，会造成巨大的资源开销；对于数据买家来说，分析和获取数据需要花费大量资金[5]；全部数据中只有一小部分具有分析价值；在某种程度上还会增加数据的误差，分析起来存在一定难度。

### 2.2.2 论坛文本数据分析

论坛，亦称讨论区。通过分析论坛数据的，能够提取关键词、分析用户和关键词之间的关联，以及建立 SLN图论模型。正如前文所介绍，SLN的论坛通常附在主题课程后，形式和功能比较典型和简单。

图2-2是中文MOOC平台学堂在线上某课程的讨论区，拥有鲜明的三级结构:发帖-跟帖-回复。左侧是发帖列表，每一个发帖、跟帖、回复都可以看作是一次针对某个关键词的发言，也就是用户和关键词之间的一次关联。一个帖子的跟帖和回复越多，说明他的观点越具有影响力。用户可以对已有的内容，包括发帖和回帖进行投票，可以投赞成票或反对票。

在某个SLN的论坛中，在对论坛文本数据进行挖掘之前，应该首先利用文本分析技术和相关处理工具把杂乱无章的论坛数据进行清洗、归类、筛选、整理，然后进行初步的文本分析程序，然后将这些过程的结果作为输入数据送入文本数据挖掘过程中。这个过程即为文本数据挖掘的预处理过程。

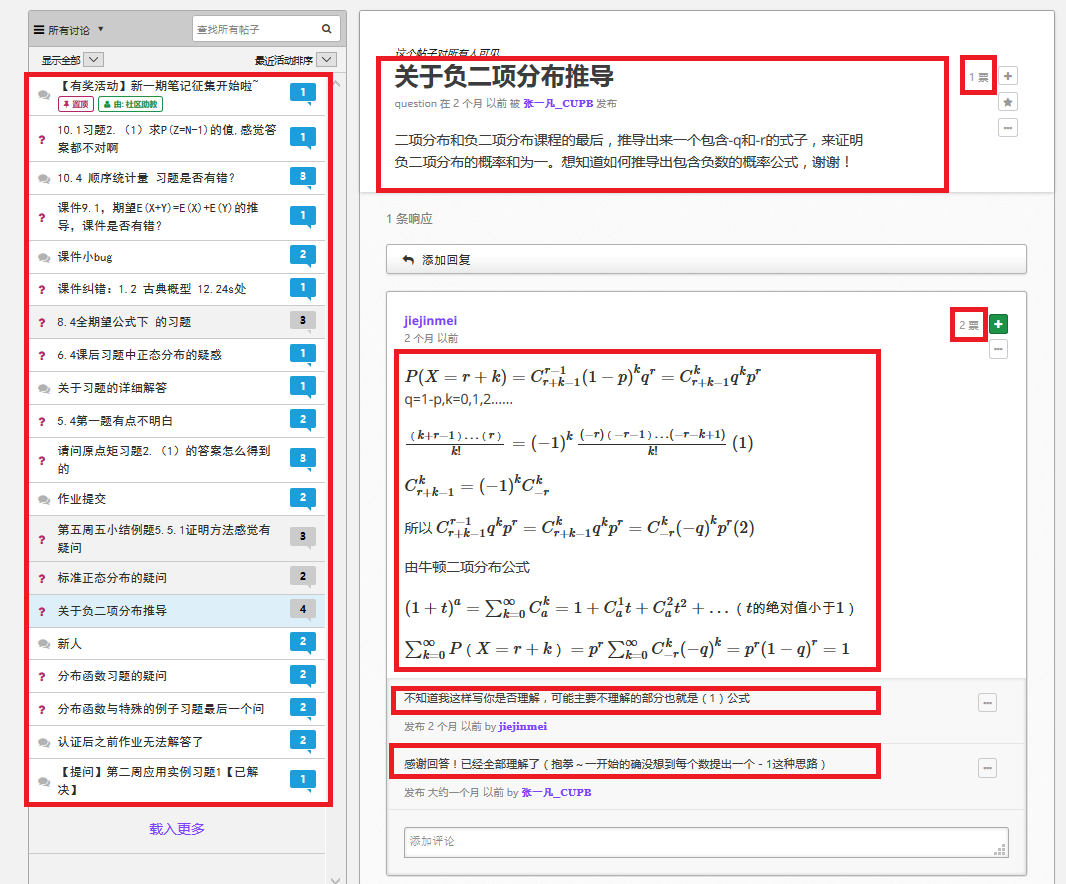


图 2-2 讨论区结构实例

图2-3是数据在进行挖掘之前需要进行的预处理过程。

文本数据挖掘过程主要进行关键词的关联分析，来获得关键词之间的潜在联系。这一部分将在第三章3.3节详细介绍。

图 2-3 数据预处理流程

输入

**数据预处理**

数据挖掘

**清洗数据**

·提取纯文本

·统一编码格式

**提纯数据**

·去除无效符号

·去除错误信息

**·**建立用户词典

·去除停用词

**文本分析**

·分词处理

·提取关键词

**·**计算权重

### 2.2.3 论坛数据优势

论坛数据的优势在于，发帖者可以对管理跟帖，以保证其内容符合预期；用户通过自觉自愿的方式来维护论坛内容的质量。论坛管理员或该讨论区所属课程的老师、助教也可以管理帖子。这样一来，论坛数据的质量就比点击流数据高很多。

此外，论坛数据是开放的，容易获取；相比之下，点击流数据比较依赖专业网站管理工具。

综合考虑，选择分析论坛文本数据。

## 2.3 SLN图论模型

### 2.3.1 图论模型的节点和边

SLN的网络模型可用图论模型来构造。图中的节点（node）代表用户和关键词，图中的边（link）代表用户和关键词之间的联系。边的权值用来表示节点之间相互关联的密切程度。图2-3表示根据论坛数据构建的SLN图论模型。简单起见，SLN的图论模型的边均为无向边。

图 2-4 根据论坛数据建立模型

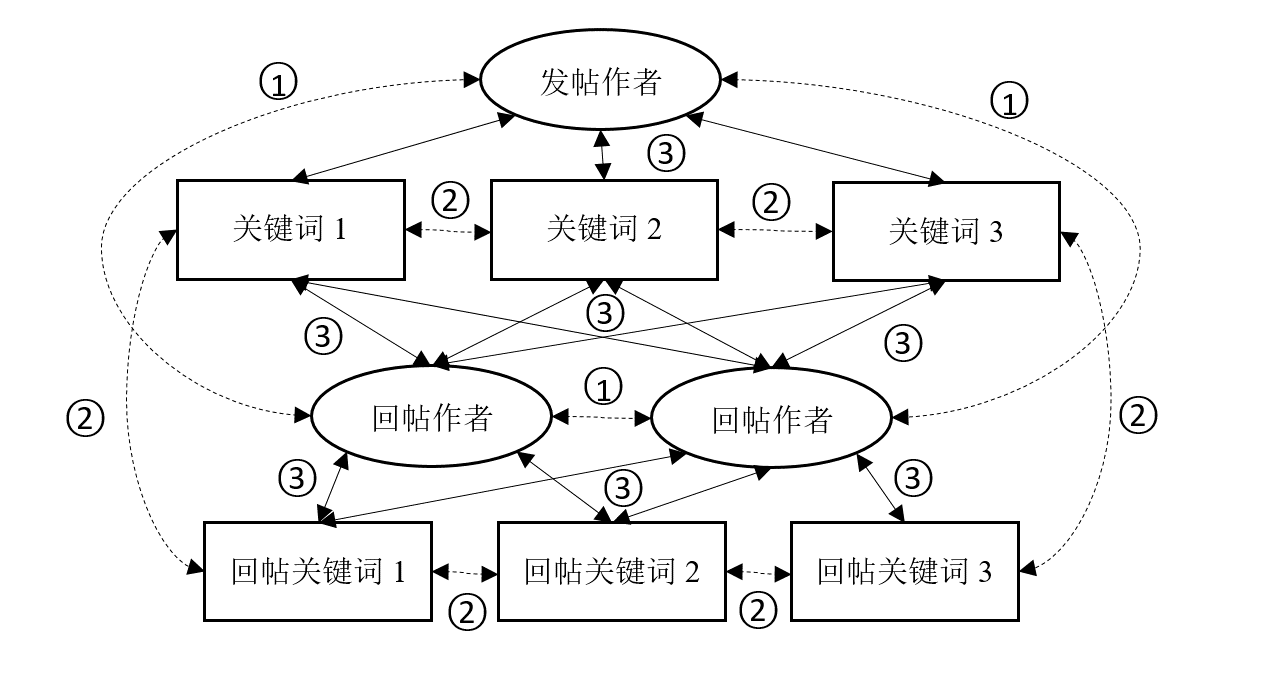


图2-4中，椭圆代表用户，矩形代表关键词。虚线边表示通过相似度和关联度建立的边，实线边代表通过用户的使用建立的边。具体来说，边分为以下三种：

#### （1）用户之间的边

该模型通过边来表示两个用户之间的相似程度。衡量相似程度的是一组描述用户的属性，如年龄、地理位置、教育背景、同时选择了某课程等。式（2-2）给出了节点和之间连一条边的所需的条件[2]。

 式 （2-2）

说明：如果和同时包含特质，那么，反之为0；为全部特质的集合。这个公式的意义是:若两个用户有至少个属性相同，那么在SLN的图论模型中这两个节点之间就会有一条边。和之间的相似度越高，这条边的权值越大。

#### （2）关键词之间的边

由关键词所指代的信息和知识是SLN图论模型中的另一种节点。关键词能够以多种方式提取出来，如利用文本分析技术从讨论区的数据中提取关键词。定义在关键词之间的边代表在数据库中关键词之间的关联程度。这个关联程度由数据挖掘技术的核心技术之一——关联分析来决定的。通过支持度-置信度框架，能够计算出关键词之间的关联程度；关联程度越高，这条边的权值就越高。

在3.3节中，将详细介绍关联分析以及如何计算权值。

#### （3）用户和关键词之间的边

通过发帖和跟帖，知识和信息在用户之间流动。通过观察，具体规则为：对每一个帖子，无论发帖或回帖，建立一条帖子作者和帖子关键词之间的边；对每一个回帖，建立一条回帖作者和发帖关键词的边。帖子的赞成/反对数越高，该帖的权重越大。特别的，在评价一篇帖子的过程中，赞成票和反对票都是重要的依据，不能只看赞成和反对的差值。一篇有1000人赞成，1200人反对的帖子，他的影响力远远超过一篇有100人赞成，50人反对的帖子。

### 2.3.2 图论模型算法

在SLN图论模型上，我们要利用许多图论算法对其进行处理和分析。如搜索算法、最短路算法、查找直径、查找中心点。图2-5是讨论区图论模型的一个例子。其中圆形代表用户，矩形代表关键词；边越长，表示节点之间距离越远。应注意，一条边的权值与其所表示的距离成负相关，权值越大，距离越近；反之亦然。

1

3

2

图 2-5 图论模型示例

#### 搜索算法

搜索算法是一种基础的图论算法，即根据一定的规则对图中所有的结点进行遍历，对满足条件的节点进行操作。图的搜索算法分为深度优先搜索（DFS）和广度优先搜索（BFS）[10]。对于我们的模型来说，采用DFS或BFS均可。

设置一个距离阈值，搜索时，将队列头的节点作为当前结点，对与当前结点以一条边直接相连且与当前结点不在同一组的另一节点来说，如果这条边的长度小于距离阈值，则将该节点加入当前节点所在的组，并加到队列尾，直至所有节点都被拓展。

该处理方式应用到图2-5所示的模型上得到图2-6所示的结果。搜索算法是聚类分析过程的第一步，其目的是将复杂的图通过简单有效的规则初步分成K组，组内部节点之间的距离相对组外的节点更近。所得的K即后续K均值方法的参数。

图 2-6 图论模型示例

1

3

2

#### 最短路算法

在SLN图论模型的处理中，所求的最短路都是单源最短路算法。使用Dijkstra算法来处理[10]。

#### 查找质心

在聚类分析时，我们使用了K均值算法。在K均值算法中需要寻找一个临时簇的质心。对SLN图论模型来说，簇的质心即连通子图的中心节点。关于查找图的中心节点

**定义** 图的中心节点连通图中与其他节点的最短距离之和最近的节点。

**定义** 图的直径 图中任意两个点的最短距离中最大那一对点之间的最短路径，直径可能有多条。为解决这个问题，引入两个定理[11]：

**定理1** 连通图上距离某一点距离最远的点，必定是直径的端点。

通过应用定理1设计算法，在图中随机指定某一个节点，通过1次计算单源最短路过程，找出离它最远的节点；然后以这个节点为源点，再通过1次计算单源最短路过程，找出离它最远的节点，这两个节点之间的最短路径，即为图的直径。

**定理2** 连通图的质心一定在图的某一条直径上。

通过应用定理2设计算法，在直径上的节点内，找出使距离和最小的节点即可。

将两个定理的算法结合起来，在点集所代表的子图中找出一条直径，在这条直径上找出中心点，即为簇的质心。在图2-6中，包含{A，E，F，1，2}五个节点的簇所构成的子图的直径和质心如图2-7所示，其中实线表示直径，灰色矩形节点表示质心。

图 2-7 直径和质心示例

1

2

找到所有簇的质心后，图中全部节点再次选择离自己最近的质心；选择了同一个质心的节点分为一个簇，再次经过上述过程，直到绝大部分或全部的节点所属的簇不变为止。具体算法将在3.4节聚类分析中详细介绍。

# 第三章 数据挖掘技术

## 3.1 数据挖掘技术简介

互联网以及计算机技术的飞速发展使得数据收集和存储技术相应得到极大进步，各组织机构、企业和学校甚至个人都能积累巨量的数据。传统的数据分析工具和技术已经无法处理这个量级的数据；或者非传统形式的数据，即使体量不大，也无法用传统技术来获取有用信息。

数据挖掘技术是指将传统数据分析方法和处理大量数据的算法相结合的一种数据处理技术。借助数据挖掘技术，我们能够应对分析非传统类型数据、分析海量数据以及从已有数据中发现新信息等挑战。

数据挖掘的主要任务是描述和预测。描述任务是找出数据中的潜在联系；预测任务是根据已有数据推测特定属性的数据。数据挖掘中最常用的方法是聚类分析和关联分析。

在利用数据挖掘技术获得信息之前，首先要获取足量的数据；其次，对收集到的数据进行预处理。之后，再对其进行关联分析和聚类分析。

本章后续小节详细介绍这些步骤。3.2节介绍网络爬虫技术；3.3节介绍数据预处理；3.4节介绍文本数据的关联分析；3.4节介绍聚类分析和K均值算法在SLN图论模型上的应用。

## 3.2 数据收集:网络爬虫技术

网络爬虫是自动提取网页内容的程序。图3-1给出了网络爬虫基本算法的伪代码。其具体流程是：

1. 利用浏览器的调试功能或其他方式和工具跟踪数据来源，找到获得数据的HTTP（或AJAX、JSONP）请求后，加入HTTP请求集{H};
2. 选择需要获取的字段和获取过则，建立字段-规则集{DR};
3. 从HTTP请求集中取出一条HTTP请求，发送后接收响应;
4. 根据一定的规则（如正则表达式）规定的路径获取目标字段数据；
5. 将获取的所有字段的数据封装为一个对象保存到输出文件或输出到控制台（屏幕）；
6. 重复(2)~(4)直到取完HTTP请求集或满足其他停止条件。

网络爬虫是搜索引擎的重要组成部分。大量、长时间、重复地使用网络爬虫将会有巨大的资源开销，影响被爬取网站服务器的性能。因此绝大部分网站都会使用登录验证、拒绝同一IP、同一浏览器、相同时间间隔的访问来使爬虫失效。解决方案：

图 3-1 网络爬虫基本算法

1: establish {H} # 建立HTTP请求集{H}

2: establish {DR} # 建立字段-规则集

3: FOR request : {H} # 取出HTTP请求集{H}中的一条记录

4: response = HTTPGET request # 获得请求的响应

5: FOR data, rule : {DR} # 对字段集中每一个字段进行处理

6: item->data = PARSE response, rule # 根据rule规则从响应中取出data字段

7: END FOR

8: OUTPUT item ; # 输出或保存item数据

9: IF quit\_condition = TRUE # 满足停止条件

10: EXIT

11: END IF

12: END FOR

1. 登录验证：使用HTTP POST方法提交登录表单，并在HTTP请求的Header中加入Cookies保持登录状态；
2. 拒绝同一IP：建立IP池随机访问；
3. 拒绝同一浏览器：随机更换每次请求的浏览器；
4. 拒绝同一间隔时间：设置每次的访问间隔在一定范围内随机，如0.5s~1.5s，既保证随机的幅度避免被服务器屏蔽，又兼顾程序执行效率。

网络爬虫技术是我们能够用自定义字段集获取目标网页上的数据集。表3-1给出了SLN图论模型网络所需的字段。

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 必须 | 说明 |
| Pin | String | 必须 | 用户标识 |
| Content | String | 必须 | 帖子内容 |
| Type | String | 必须 | 发帖或回帖 |
| Vote | Integer | 可选 | 投票数 |
| Quote | Integer | 可选 | 引用数 |

表 3-1 论坛数据分析字段

经过网络爬虫程序获取的数据是原始粗糙的，它的类型、格式、编码不统一，还不能作进一步的分析。在进行网络建模之前，需要把数据仓库中所有的帖子内容作预处理，才能进行文本分析和建模。

## 3.3 数据预处理

通过网络爬虫算法收集的数据转化为表3-1的形式后，称为原始数据。在进行建模和分析之前需要对*content*字段的数据进行预处理，即通过如图2-3的流程，进行数据清洗、数据提纯和文本分析。

### 3.2.1 数据清洗过程

数据清洗是预处理的第一步。

首先，应在原始数据中提取纯文本。在收集的数据中不仅有纯文本，还有一些图片、公式、表情等，还有一些无法被编码的数据，应该一概去除。

其次，应该统一纯文本数据的编码格式。提取出纯文本后，根据网站的设置，数据一般统一编码为GBK格式或UTF-8格式。考虑到重点分析中文MOOC平台，GBK格式下的中文显示为乱码（部分过时的编辑器或IDE会将GBK格式下的中文转换为正常文字显示，此时UTF-8格式的中文显示可能反而是乱码）。目前中文WEB站点基本使用UTF-8格式，所以最好编码为UTF-8格式。

### 3.2.2 数据提纯过程

数据提纯过程将统一编码（UTF-8）的纯文本数据变为便于分析的有效数据。

第一步，应去除无效字符。无效字符是指无语义或无分析价值的字符。在本文中，假定所有英文字母，数字，中英文标点符号、表情符号均为无效字符。中文数字按此规则也应去除，但实际情况中，中文数字可以与其他单子组词，含义也会超过其本身的意义，且包含同一数字的词汇重复率极低，所以不用特殊处理。

第二步，去除数据中的错误信息。在许多网站上发帖、回帖可以使用多媒体资源或其他第三方资源。当这些资源的载入或显示发生错误时，网站服务器会自动使用缺省的错误信息替换，如“资源载入错误”“视频不存在”等。而且这些错误信息每个网站都有可能不一样，需要将这样的字符串同时加入用户词典和停用词典，以确保它们作为一个整体被剔除。去除无效字符、错误信息后的统一编码格式纯文本数据称为高质量数据。

第三步，通过先验方式建立部分的用户词典和停用词典，并去除停用词。

用户词典是一类专有名词的词典，在普通文本处理程序中无法自动识别。考虑如下一句话：“新华社长江前线报道”，如果把“新华社”加入用户词典，则分词结果为“新华社/长江/前线/报道”；如果把“江前线”加入用户词典，则分词结果为“新华/社长/江前线/报道”。为了文本分析程序达到预期，大部分人名、定理、专业名词等应加入用户词典。

停用词是用户设定的、在目标文本数据中予以删除或忽略的单词，这些单词往往不带语义而又有一定概率出现，如各类语气词、象声词、虚词等。甚至比某些关键信息出现次数还多。它们对文本分析来说是有害的。

去除停用词的高质量数据称为有效数据，将被分词和提取关键词。

### 3.2.3 文本分析程序

有效数据是理想的文本分析程序的输入数据，在完整保留原始数据的语义信息的同时尽量剔除垃圾数据。文本分析程序有两个功能:

1. 分词：把文本数据通过分词方案分成单词；分词后，所有的词语构成一个分词词典，用来建立SLN图论模型。
2. 提取关键词：通过算法计算每个单词的权重，由大到小排列。较为著名的关键词加权算法有TF-IDF，textrank等。

图3-2给出了文本预处理的步骤。经过大量数据分析，用户词典中平均每个单词大小为4.25字节。在经过文本分析后，分词词典(包含单词权重)和原始数据中除content字段外的其他数据一起被送入之后的图论建模和数据挖掘过程。

数据源

∞

原始数据

高质量数据

有效数据

用户词典

网络爬虫

去除无效字符

去除错误信息

用户词典

停用词典

分词

提取关键词

图 3-2 数据预处理步骤

## 3.4 关联分析

关联分析的目标是研究一个项集的出现是否对另一个项集的出现产生影响。一般说来，出现比不出现更具代表性，所以仅讨论两个项集是否同时出现的关系。

设分别表示两个项集，每个项集中包含若干个项；某个项集出现，等价于项集里所包含的所有项都出现。项集在全部事件里出现的次数，称为“支持度计数”。形如的蕴含表达式称为“关联规则”。关联规则的强度可以由“支持度”和“置信度”来衡量。支持度说明项集本身出现的频繁程度，置信度说明在包含的事件中出现的频繁程度[1]。

支持度即支持度计数（即同时出现的次数）与事件总数的比值。置信度是的支持度计数与支持度计数的比值。

由支持度和置信度构建的规则集称为支持度-置信度框架。

支持度是一个重要的指标，支持度低的规则一般都是偶然出现或无意义的。忽略去除支持度低的规则能帮我们过滤。同样的，置信度低的规则说明两个事件出现的相关程度低，该规则不可靠。特别的，关联分析的结果不一定能用简单的因果关系来解释，有可能是长期的、多层次、多方面的内因和外因综合作用的结果。一方面要充分利用关联分析的结果来指导决策，另一方面要深入地研究这些作用的来源。

例如，在一个电脑外设店中某天有10名客人，有5名买了键盘，8名买了鼠标，3人既买了键盘又买了鼠标，那么规则的支持度为3/10=0.3，置信度为3/5=0.6。另外的，规则的置信度为3/8=0.375，所以在既买了鼠标又买了键盘的客户中，“键盘的销售带动了鼠标的销售”的说法比“鼠标的销售带动了键盘的销售”的说法更具可信度。

在SLN的大数据挖掘中，关联分析的任务就是找出所有支持度大于等于*minsup*（支持度阈值）及置信度大于等于*minconf*（置信度阈值）的关联规则。这一过程称为关联规则的发现。在上一节数据预处理所输出的有效数据中，原始数据中每出现5.52个词时，用户词典中就会新增一个词；即每5.52\*4.25字节=23.46字节数据就会出现一个新词。那么当面对23.46兆字节的文本数据时，就会产生106级别的新词。所以在产生新规则时，引入和支持度阈值和置信度阈值能够极大的减少系统的开销和资源的浪费。

支持度计数大于等于支持度阈值所对应的支持度计数的项集被称为频繁项集，简称繁项集。

关联规则的发现算法基于如下原理:如果一个项集是繁项集，那么它的所有子集都是繁项集；反之，如果一个项集不是繁项集，那么它的所有超集都不是繁项集。这个原理被称为“先验（*Apriori*）原理”。

根据先验原理，我们以支持度计数大于等于支持度阈值的单项集（又称1-项集，即仅包含1个项的项集）为出发点，通过迭代的方法，在保证支持度的情况下不断扩充项集，直到(1)扩充到指定大小或最大；（2）无支持度大于等于阈值的项可扩展。这个算法即*Aprori*算法，其伪代码由程序清单3-2给出[1]。

图 3-2 *Apriori*算法

1: size=1 #项集大小, {V}表示点集

2: {F*size*}={ {v} | v{V} & sup(v)minsup } #sup()为项或项集的支持度

3: REPEAT

4: *size*=*size*+1

5: FOR {F*i*} : {F*size*-1} #扫描每个项集

6: FOR x : {V} #扫描每个项

7: IF sup({F*i*v})minsup #支持度条件成立

8: { F*i*v}{F*size*} #扩展繁项集加入

9: END IF

10: END FOR

11: END FOR

12: UNTIL quit\_condition=TRUE #退出条件成立时停止算法

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| minsup | 单词数 | 1-项集数 | 2-项集数 | 3-项集数 | 4-项集数 |
| 0.01 | 1011 | 217 | 116 | 24 | 6 |
| 0.02 | 1011 | 92 | 13 | 1 | 0 |
| 0.03 | 1011 | 46 | 3 | 0 | 0 |

在*Apriori*算法产生的繁项集中我们可以提取出形如的关联规则，其中称为“规则前件”，称为“规则后件”。这是关联分析最重要的一步。由于自然语言文本数据和SLN的特殊性，这一步比其他数据挖掘过程要简化很多。

表 3-2 项目集数

在某MOOC平台论坛中，通过本章前述小节处理后的有效数据经前述文本分析和*Apriori*算法产生的繁项集规模如表3-2和图3-3所示。

图 3-3 项目集数示意图

在支持度阈值大于0.03时，3-项集数满足条件的个数为0，容易丢弃隐含的有用信息。所以支持度阈值区间在0.01~0.03之间。此时，4项集的个数最多不过6个，高复杂度分析的次数非常少，所以优化的空间并不大，可省略。

提取规则时，可以遵循如下方法：枚举每个项集，分解为和两个互补的子项集，考虑规则和的置信度是否满足置信度阈值。一个项集的分解方法有种，4项集有14种，3项集有6种，2项集只有2种。

例如，当支持度度阈值为0.02时，上述3-项集为{证书，申请，截止}，支持度为0.023923，分解方法和支持度如表3-4所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 规则前件 | 证书,申请 | 证书,截止 | 申请，截止 | 证书 | 申请 | 截止 |
| 规则后件 | 截止 | 申请 | 证书 | 申请，截止 | 证书,截止 | 证书,申请 |
| 置信度 | 100.0% | 100.0% | 100.0% | 27.7% | 21.7% | 100.0% |

表 3-4 规则提取

由此可见，当学生们谈论到“证书”的话题时，他们不一定会同时谈到“申请”和“截止”，也许会讨论些什么别的。反之，如果讨论到“证书”和“申请”时，他们肯定会讨论到“截止”的问题。由于置信度太低而项集太大，结果中出现了很多语义上的搭配短语，如“申请证书”“截止申请”“证书申请截止”等。所以，有必要提高支持度阈值而缩小项集的大小。

当支持度阈值为0.04是，上述2-项集为{认证，申请}和{微积分，修课}，支持度均为0.043062，分解方法和支持度如表3-5所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 规则前件 | 认证 | 申请 | 微积分 | 修课 |
| 规则后件 | 申请 | 认证 | 修课 | 微积分 |
| 置信度 | 39.1% | 39.1% | 31.0% | 75.0% |

表 3-5 规则提取

由此可见，在支持度阈值较高的情况下分析2-项集时的结论为：当谈论到“认证”和“申请”时，互相有大概39.1%的概率提及对方；当谈论到修课时，学生们有75%的概率会说到“微积分”，而说到“微积分”时，有大概69.0%的时候并没有谈论到“修课”。

在SLN图论建模中，节点的权重与项的支持度正相关；节点之间边的权重与规则置信度正相关。

## 3.5 聚类分析

### 3.5.1 聚类分析简介

聚类分析的任务是把数据分成有意义的簇或组，一定程度上类似分类、划分等传统数据分析概念。组内元素之间的距离越近、越相似，组间元素之间的距离越远、相差越大，聚类的效果越好。聚类有不同的类型，应根据SLN的大数据挖掘的具体任务需求来选择。聚类分类：层次和划分的；互斥、模糊和重叠的；以及完全和部分的[1]。

层次聚类，即允许簇有子簇的树形聚类；划分聚类，即不允许簇有子簇的平行聚类。在SLN中，由于课程、话题之间具有天然的层次性，应该使用层次聚类。

互斥聚类，即每个项分配到单个簇；重叠聚类，即一个项可以同时属于多个簇；模糊聚类是重叠聚类的一种扩展，即每个项可以以一个[0,1]的系数属于一个簇。通常来说，规定对于每一个项，这些概率的和为1。实际操作的过程中，如果某个项必须要分配给一个簇，那么就选择具有最高系数的簇，转化为互斥聚类。

完全聚类，即将每个项都指派到一个簇；部分聚类，即允许一部分项无法明确聚类或不作聚类。这些项可能代表噪声、无意义项、离群项或与本任务无关的项。这些项在文本预处理和文本分析程序中已经尽量去除，但是依然可能有残留。在本文介绍的聚类分析中，可以把这些项与已有簇的概念或定义进行比对，如果都不符合，那么把他分到一个“垃圾箱”簇中，就把部分聚类转化成了完全聚类。

根据我们任务的具体需求，应使用划分的、模糊的、部分的聚类。这种聚类较难操作和实现。通过我们的分析，应转换为划分的、互斥的、完全的聚类。

### 3.5.2 K均值算法

K均值算法是最基础的聚类算法之一。K均值算法的基础流程是[1]：

1. 选择K个初始质心；
2. 每个点选择离最近的质心；
3. 选择了同一个质心的点标记为一个簇；
4. 更新每个簇的质心
5. 重复(2)~(4)，直到质心不发生或到仅有很少一部分点可能会改变质心为止。

本节将K均值算法与SLN图论模型结合，所做的工作有确定K值和初始质心、查找最近的质心、确定点集的质心。利用2.3.2节中介绍的搜索算法、最短路算法和查找质心的算法来完成上述工作。

#### 确定K值和初始质心

K值代表了聚类分析的强度。我们希望聚类的结果是:两个数据节点的距离近到某个程度时分为一个簇，距离远到某个程度时就自动脱离。如果K太大，会把本该属于一个簇的项分隔开；K太小，会把距离较远的项集聚类在一起。

利用搜索算法，首先将整个图标记为*k*个簇，在保持基本稳定的情况下，一个距离阈值对应一个标记结果。将某个阈值对应的结果显示出来后，可以根据实际情况调整K值，然后重新执行搜索算法。在一个相对合理的搜索结果和K值下，可以丢弃掉一部分只包含单项的、离群的、无意义的簇。

在我们用搜索算法获得初始的簇后，可以在每个簇内随机选定一个点，或选用初始簇内的质心。关于如何在一个点集内确定质心的算法将在后面进行详细介绍。

#### 查找最近质心

确定好质心后，每个项查找离自己最近的质心。不妨把问题转化一下，使用Dijkstra算法计算每个质心到图中其他点的最短距离，在此过程中，每个点记录离自己最近的质心的编号和距离值。这个过程称为renew，图3-3给出该过程的伪代码。

图 3-3 renew过程

1: FOR *i*: 1 TO N

2: H*i*=null # 初始化每个节点所属的质心

3: D*i*=∞ # 初始化每个节点到质心的距离

4: END FOR

5: FOR heart: {Heart} # 枚举质心集

6: Dijkstra(heart)→{Dis*i* | *i*=1,2,…,N} # 计算质心到每个节点的最短距离

7: FOR *i*: 1 TO N

8: IF D*i*>Dis*i* THEN # 更新维护值

9: D*i*←Dis*i*

10: H*i*←heart

11: END IF

12: END FOR

13: END FOR

#### 查找点集的质心

根据点集重新选择质心，就是在点集所对应的子图中找到中心点。图的中心点到图中每个点的最短距离之和最小，与点集质心的概念重合。首先通过两次Dijkstra算法查找直径，然后再直径上枚举每个节点，找到使距离和最小的节点。该过程称为barycenter。设子图的点集为{V}，质心集为{Heart}，查找图中心点的算法如图3-4所示。

FOR *v* : {V} #遍历点集分初始簇和确定K，K初始为0

IF cluster(*v*) not existed THEN

DFS(*v*);

K=K+1

END IF

END FOR

REPEAT

renew {V} #更新每个节点的簇

{Heart}←barycenter #更新质心集

UNTIL quit\_condition=TRUE #直到满足停止条件

图 3-5 SLN图论模型K均值算法

图 3-4 barycenter过程查找图中心点

1: {D}←diameter {V} # 通过两次Dijkstra算法获得直径

2: FOR *d* : {D} # 按顺序遍历直径上的每个点

3: IF sum\_of\_distance( *d*→{V} ) < min\_sod # 找最小距离和的点

4: mid\_sod←sum\_of\_distance ( *d*→{V} )

5: heart←*d* # 更新质心

6: END IF

7: END FOR

综和本节算法，SLN图论模型的K均值算法过程如图3-5所示。

# 第四章 数据挖掘成果

本章通过对学堂在线（www.xuetangx.com）上的热门课程“C++程序设计基础”的讨论区进行SLN图论建模和数据挖掘，并展示各个步骤中所做的工作以及结果，分析和应用数据挖掘成果。

## 4.1 获取数据

本文获取数据的工作基于Python2.7 Scrapy3网络爬虫框架。观察目标课程讨论区某，发现用户讨论比较积极热烈，但几乎没有投票，这说明投票是偶然行为并不能代表该用户群体的行为特征。故取帖子作者，帖子类型和帖子内容三个字段来建立原始数据。表4-1给出了部分原始数据，仅供参考。

|  |  |  |
| --- | --- | --- |
| **用户名** | **类型** | **帖子内容** |
| 绿领巾童鞋 | 发帖 | 同学们好，新一期笔记征集活动又开始啦…… |
| kdd | 发帖 | 我们会定期发一些C++相关的文章,欢迎关注…… |
| kdd | 发帖 | 课程官方QQ群…… |
| Houwei1975 | 跟帖 | 加了2次还是没通过啊？ |
| Lucky\_Boy | 发帖 | 肯定不会考零分啊？讨论区里不只我一个人反映难道是…… |
| 想飞的鸵鸟258564 | 跟帖 | 就是啊，我也是零分。都交了钱了。 |
| 玄蓂 | 跟帖 | 我的也是零分啊，显示答案对过了，不是零分啊？什么情况？…… |
| ayyzlpy | 跟帖 | 我也是零分 |
| qq3190387232 | 跟帖 | 我虽然学的不好，但也不至于考零分，难道错题倒扣分？…… |
| 踞山瞰海 | 发帖 | 若用户输入的值超出了预设的范围，我们可以用如下语句…… |
| 踞山瞰海 | 发帖 | 这两个命令都能够将别的类型转化为int型，它们有什么区别呢？ |
| 踞山瞰海 | 发帖 | 一、我用的是累加法，变量ans用int型的话…… |
| 闲人242545 | 跟帖 | 用long long类型就不会溢出，也不会有误差。 |
| blexxx | 跟帖 | 我的理解，double类型初始化的时候比如…… |
| 踞山瞰海 | 发帖 | 根据该式，分子每次都增大x²倍，那么如果x越小，…… |
| gilvaa | 发帖 | 请大神们帮忙看一下，这个为什么只有60分？ |
| blexxx | 跟帖 | 示例输出里没逗号 |
| 学堂活动小队长 | 发帖 | 上了这么久的课，汲取了这么多知识，大家是否想…… |
| 天霸我来也 | 跟帖 | 好 |
| 学堂活动小队长 | 发帖 | 学堂君给大家送年末福利啦~！现在到学堂广场发帖晒课程笔记…… |
| …… | …… | …… |

表 4-1 部分原始数据

## 4.2 预处理

预处理过程将混杂格式的数据变为统一形式、统一编码格式的纯文本数据。将这些纯文本字符串中的无效、无用、无价值字符去掉，如英文字母、符号、公式、图片等。如有些WEB站点会在无法载入图片、表情、视频等多媒体插件时显示“载入错误”等信息，将有可能出现很多次，应将其删除。将专有名词加入到用户字典中，比如“柯西不等式”等，以免其不符合中文语法而被错误地分词。

表4-2给出了部分用户词典以及部分停用词。

|  |  |  |  |
| --- | --- | --- | --- |
| **用户词典** | **停用词** | | |
| 学堂在线 | **中文词汇** | **ASCII码表字符** | **其他字符** |
| 载入错误 | 载入错误 | 0~9 | 。 |
| 张广远 | 本 | A~Z | ， |
| 韩云端 | 将 | a~z | ！ |
| 郑莉 | 在 | ! | ？ |
| 有难度 | 时 | @ | 、 |
| 花钱 | 些 | # | ： |
| 工科生 | 得 | $ | ； |
| 理科生 | 里 | % | ∩ |
| 文科生 | 外 | ^ | 《 |
| 一道题 | 出来 | & | 》 |
| 正无穷 | 进去 | \* | ～ |
| 负无穷 | 但 | ( | … |
| 拉格朗日 | 但是 | ) | π |
| 中值定理 | 的 | . | → |
| …… | …… | …… | …… |

表 4-2 部分用户词典和停用词

表4-3给出了部分在经过清洗、提纯等文本预处理后所得的回帖内容。

|  |
| --- |
| **帖子内容** |
| 同学们好新一期笔记征集活动又开始啦秀出你的笔记…… |
| 我们会定期发一些相关的文章欢迎关注 |
| 课程官方群 |
| 加了次还是没通过啊 |
| 肯定不会考零分啊讨论区里不只我一个人反映难道是系统错误…… |
| 就是啊我也是零分都交了钱了 |
| 我的也是零分啊显示答案对过了不是零分啊什么情况 |
| 我也是零分 |
| 我虽然学的不好但也不至于考零分难道错题倒扣分…… |
| 若用户输入的值超出了预设的范围我们可以用如下语句令其重新输入…… |
| 这两个命令都能够将别的类型转化为型它们有什么区别呢 |
| 我用的是累加法变量用型的话较大的结果好像会超出的范围造成错误…… |
| 用类型就不会溢出也不会有误差 |
| 我的理解类型初始化的时候比如一定大的时候不断累加这个小数部分后…… |
| 根据该式分子每次都增大倍那么如果越小则整个分式就越早开始缩小…… |
| …… |

表 4-3 文本预处理结果

经过文本预处理后，去掉的无用数据占比达20.6%，不仅减少了系统开销和资源浪费，还将大大降低了无用信息所造成的负面影响。

## 4.3 分词及提取关键词

将文本分析所得的结果送入分词、提取关键词、建立网络模型等算法的输入端，将这些算法的输出结果作为预处理程序的输出结果等待数据挖掘。

本文采用优秀的开源分词方案JIEBA分词，基于Python2.7平台开发分词和提取关键词程序，并设置权值系数来调整关键词权值的整体大小。JIEBA分词有两种方案，一种是完全版，一种是快速版。完整版将输出目标文本中每一小段的所有分词方案，快速版直接将文本划分成单词。本文我们选取快速分词，表4-4给出了经过JIEBA分词的部分数据。

|  |
| --- |
| **帖子内容的单词** |
| 新一期 笔记 征集 活动 开始 秀出 笔记 机会 获得 元 认证 优惠券 呼朋唤友赞 最新版 德国 钢笔 精美 铁质 书签 详细 内容 查看 |
| 定期 文章 欢迎 关注 |
| 官方 群 |
| 加 通过 |
| 肯定 考 零分 讨论 区里 人 反映 系统 错误 多天 人 修改 官方 回应 |
| 就是 零分 交 钱 |
| 零分 显示 答案 对过 零分 情况 |
| 零分 |
| 学 不好 不至于 考 零分 错题 倒扣分 分数 扣 完 事实 显示 考试 显示 确实 零分 只想 问问 有没有 谁 显示 考试 零分 |
| 用户 输入 值 超出 预设 范围 如下 语句 令其 重新 输入 用户 输入 字母 之类 怎么 达到 相同 目的 语段 貌似 出错 |
| 命令 能够 类型 转化 为型 它们 区别 |
| 累加法 变量 用型 较大 结果 好像 超出 范围 造成 错误 于是 改用 或型 或 范围 完全 涵盖 结果 结果 较 偏差 结果 正确 结果 困惑 之前 数据 较小时 递归 法 主要 问题 重复 计算 太多 如 计算 时会 两次 有没有 可能 算出 存在 数组 下次 直接 取出 来用 试过 成功 求教 若渴 |
| 类型 溢出 误差 |
| …… |

表 4-4 分词后的文本数据

经过自动分词后，可能会有一些遗漏的专有名词或停用词，可以多次进行分词、多次调整用户词典和停用词典，使分词更加完善。分词后，应对全部数据提取关键词。在JIEBA分词中有两个算法用来提取关键词，TF-IDF[13]和TextRank[14]。本文不讨论这两种算法的详细内容，仅将两种提取关键词和权值的部分结果在图4-5中给出。

|  |  |  |  |
| --- | --- | --- | --- |
| **TextRank** | | **TF-IDF** | |
| 输入 | 0.905750282164 | 递归 | 0.0666233371875 |
| 函数 | 0.828989751315 | 输入 | 0.066282881095 |
| 输出 | 0.736773849466 | 零分 | 0.0645810353259 |
| 数据 | 0.702308584441 | 函数 | 0.0644910772928 |
| 时候 | 0.645010230026 | 输出 | 0.05933240765 |
| 结果 | 0.580745185474 | 代码 | 0.0469451544072 |
| 程序 | 0.580284485126 | 补码 | 0.0442139338656 |
| 使用 | 0.558696329341 | 作业 | 0.0436668797495 |
| 计算 | 0.535076414948 | 问题 | 0.0405935956276 |
| 递归 | 0.526526452566 | 程序 | 0.0396363957392 |
| 作业 | 0.505212529205 | 调用 | 0.0360036095887 |
| 代码 | 0.501013662541 | 数组 | 0.0356731754203 |
| 老师 | 0.475341833779 | 为什么 | 0.0345930521249 |
| 数组 | 0.470602797062 | 计算 | 0.0318544060148 |
| 文件 | 0.461712948452 | 数据 | 0.0312862599155 |
| 运行 | 0.459642026244 | 老师 | 0.0308245147807 |
| 引用 | 0.42562180417 | 题目 | 0.0307466790659 |
| 零分 | 0.374312154634 | 回文 | 0.0302776134111 |
| …… | …… | …… | …… |

表 4-5 部分提取关键词和权值结果

在实际操作程序中，选用任意一个均可。本文使用TF-IDF所得权值作为节点权值。分词所建立的用户词典中，平均每个单词的大小为4.19字节，与大量数据所得的平均值基本一致。

## 4.4 关联分析

本文采用C语言实现*Apriori*算法，构建支持度-置信度框架，并通过实践计算和调整关联权值。首先应用支持度剪枝，设置支持度阈值为0.15，所得1-项集个数为XXXX，2-项集个数为XXXX，3-项集个数为。对于每个三项集可以看出，XXXX是中文固定搭配，去除。在2-项集中应用置信度剪枝，得到最终的规则集。规则集中包括两个互相关联的项和规则置信度。如果形如和的规则都入选，则边的权值为两个规则中较高的。

表4-6给出了经过关联分析后生成的点集和边集，用户节点加粗表示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **点集** | | **边集** | | |
| 节点名 | 权值 | 端点1 | 端点2 | 权值 |
| **tsinghua\_zhangxuan14** | 0.0646203648805 | **kdd** | 函数 | 0.017241 |
| 递归 | 0.0602782574553 | **kdd** | **tsinghua\_suk14** | 0.043103 |
| 输入 | 0.0599702257526 | **tsinghua\_suk14** | **ChenweiSmile** | 0.086206 |
| **tsinghua\_suk14** | 0.0584660444157 | 通过 | 输出 | 0.021552 |
| 零分 | 0.0584304605330 | 完全 | 数组 | 0.012931 |
| 函数 | 0.0565256614962 | 递归 | 数组 | 0.017241 |
| 输出 | 0.0536817021595 | 补码 | 原码 | 0.017241 |
| **blexxx** | 0.0492345637185 | 函数 | 定义 | 0.017241 |
| 代码 | 0.0424741873208 | 函数 | **ChenweiSmile** | 0.021552 |
| 补码 | 0.0400030830213 | 函数 | 调用 | 0.038793 |
| …… | …… | …… | …… | …… |

表 4-6 SLN图论模型的部分点集和边集

## 4.5 建立图论模型

本文采用C语言实现建图算法，包括创建用户和关键词节点、创建由关联分析生成的节点和边。

图4-1给出了本课程的SLN图论模型。图中，红色的边表示用户和用户之间的关联，紫色的边表示关键词和关键词之间的关联，绿色的边表示用户和关键词之间的关联。

图中的节点数占总单词数的10.8%，说明绝大部分的节点和项集的支持度小于阈值；图中也有许多点满足支持度阈值的，但是由于不满足规则置信度阈值或与其相连的节点均不满足支持度阈值，所以度为0，予以剔除。

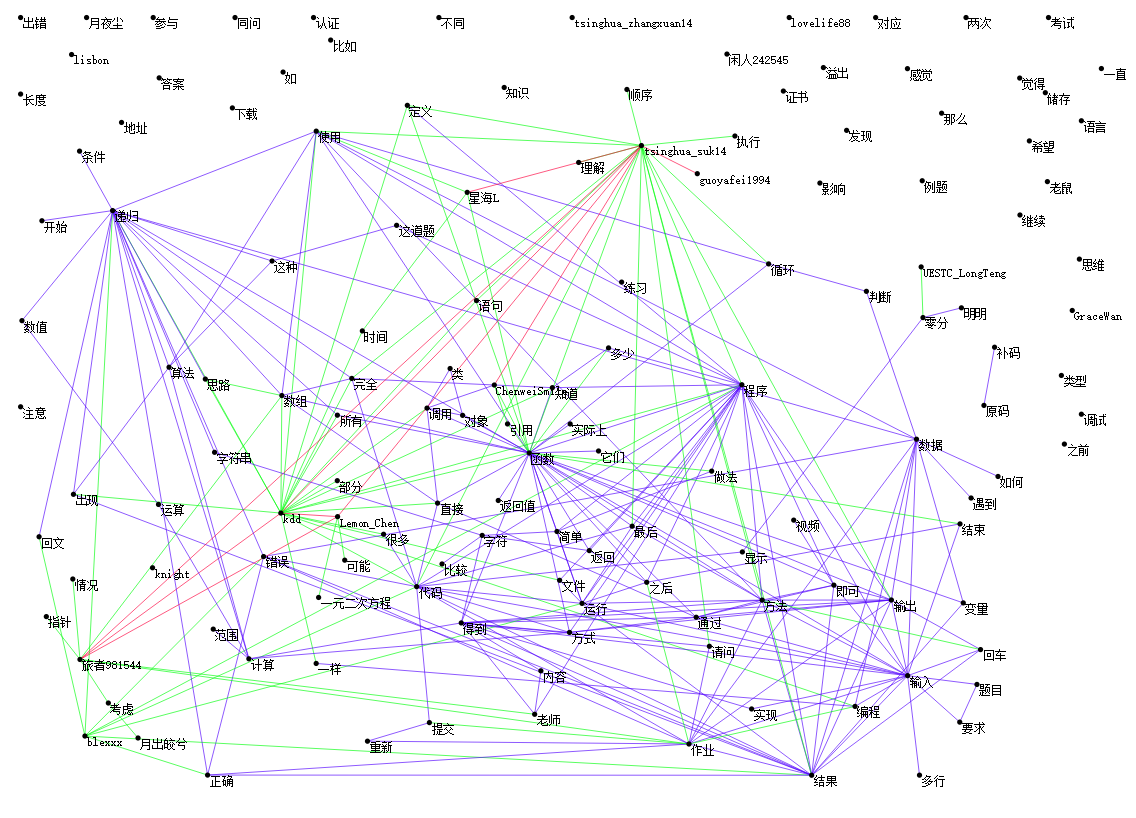
我们还可以直观地看出，虽然有些用户有多次发帖、回帖，但是由于发帖内容中符合支持度阈值的关键词太少，导致该节点的度较低。换言之，SLN图论模型能够自动识别论坛中无实际意义的发言，并剔除这些用户和信息。

图 4-1 SLN图论模型

## 4.6 聚类分析

本文采用C语言实现搜索算法、Dijkstra算法和K均值算法，将SLN图论模型分簇。在分簇时，将大小为1的簇丢弃，大于1的保留。设搜索算法距离阈值为0.5，得K=7。表4-7给出了初始聚类。初始聚类后，得出的K值和初始质心可以

|  |  |  |
| --- | --- | --- |
| **编号** | **节点名（权重）** | **质心** |
| 0 | tsinghua\_zhangxuan14(0.064620) 递归(0.060278) 条件(0.006301) | tsinghua\_zhangxuan14 |
| 1 | 输入(0.059970) 输出(0.053682) 结果(0.026733) 方法(0.016623) 回车(0.013602) 要求(0.007114) | 输入 |
| 2 | 零分(0.058430) 明明(0.009963) | 零分 |
| 3 | tsinghua\_suk14(0.058466) 函数(0.056526) 程序(0.037492) 调用(0.032575) 返回值(0.013747) 知道(0.010229) 简单(0.010123) 定义(0.009258) 对象(0.009110) 多少(0.006790) 它们(0.006445) | tsinghua\_suk14 |
| 4 | 补码(0.040003) 原码(0.017890) | 补码 |
| 5 | 计算(0.028821) 数值(0.011963) 运算(0.010571) | 计算 |
| 6 | blexxx(0.049235) 老师(0.027889) 得到(0.013051) 内容(0.006608) | blexxx |

表 4-7 初始聚类

在初始聚类和质心集上应用K均值算法，当第三次迭代完成后，所有的节点和边的簇不再改变，算法结束。表4-8给出了K均值算法的结果。

|  |  |  |
| --- | --- | --- |
| **编号** | **节点名（权重）** | **质心** |
| 0 | tsinghua\_zhangxuan14(0.064620) | tsinghua\_zhangxuan14 |
| 1 | 递归(0.060278) blexxx(0.049235) 数组(0.032276)  旅者981544(0.030772) 计算(0.028821) 回文(0.027394)  算法(0.024608) 字符串(0.024400) 正确(0.023358)  这道题(0.018463) 错误(0.016711) 指针(0.014310) 思路(0.012627) 考虑(0.012031) 数值(0.011963) 所有(0.011122) 运算(0.010571) 完全(0.009932) 范围(0.009360) 这种(0.008176) 开始(0.007918) 情况(0.007220) 条件(0.006301) 出现(0.005935) | 递归 |
| 2 | 输入(0.059970) 作业(0.039508) 数据(0.028307) 老师(0.027889) 题目(0.027818) 结果(0.026733) 编程(0.025329) 提交(0.023240) 视频(0.020739) 多行(0.017391) 方法(0.016623) 回车(0.013602) 得到(0.013051) 结束(0.012365) 即可(0.010454) 如何(0.009836) 重新(0.009537) 遇到(0.009530) 一样(0.009117) 实现(0.008427) 要求(0.007114) 内容(0.006608) | 输入 |
| 3 | tsinghua\_suk14(0.058466) 代码(0.042474) 语句(0.026779) guoyafei1994(0.021540) 一元二次方程(0.021468) 执行(0.020029) 请问(0.019488) Lemon\_Chen(0.018463) 顺序(0.011701) 理解(0.009334) 很多(0.008754) 可能(0.007718) | tsinghua\_suk14 |
| 4 | 零分(0.058430) UESTC\_LongTeng(0.030772) 显示(0.013959) 明明(0.009963) | 零分 |
| 5 | 函数(0.056526) ChenweiSmile(0.040003) kdd(0.040003)  程序(0.037492) 星海L(0.033849) 调用(0.032575) 运行(0.026391) 引用(0.025845) 文件(0.025377) 使用(0.022087) 字符(0.019604) 变量(0.018666) 之后(0.016899) 最后(0.015582) 循环(0.015486) 判断(0.014900) 返回值(0.013747) 返回(0.011973) 直接(0.011210) 做法(0.010251) 知道(0.010229) 简单(0.010123) 方式(0.009887) 比较(0.009826) 定义(0.009258) 对象(0.009110) 实际上(0.008466) 时间(0.007351) 多少(0.006790) 它们(0.006445) 部分(0.005676) | 函数 |
| 6 | 输出(0.053682) 通过(0.012462) | 输出 |

表 4-8 K均值算法结果

## 4.7 应用挖掘结果

SLN图论模型数据挖掘的目的是，用户访问包含某关键词的页面时，返回该关键词簇中权值最重的点。在网站设计时，可以人为地在页面的HTML文档中设置关键词，也可以在用户进行搜索、发帖、评论动作时对用户提交的表单内容进行获取，进行文本分析、分词等操作，找到用户所需的关键词。然后根据这些关键词所在的簇的节点进行推荐。

图4-2给出了根据K均值算法的聚类分析结果。图片使用Hypercube 1.7.0生成。

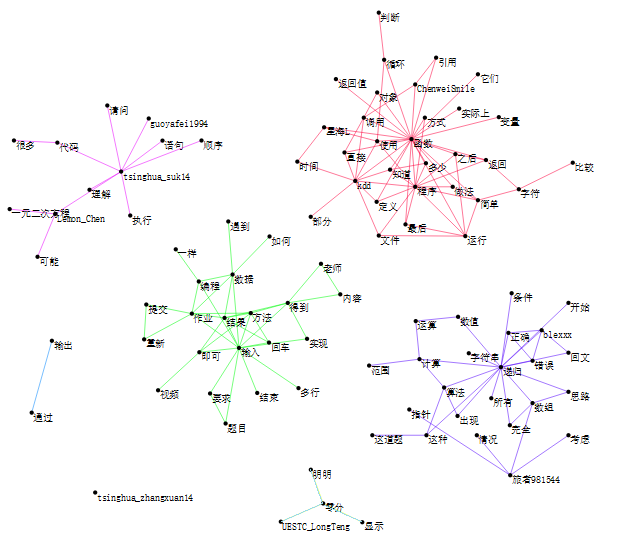
应用举例：当用户搜索“示例程序中变量数组的引用”，将其进行分词，得到关键词为： “程序”“变量”“数组”“引用”。根据聚类分析结果，该关键词属于簇1和簇5；簇1中除被查询关键词外权值最重的节点是递归(0.060278) ，blexxx(0.049235)，旅者981544(0.030772)；簇1中除被查询关键词外权值最重的节点是函数(0.056526)， ChenweiSmile(0.040003) ，kdd(0.040003) 。可向用户推送blexxx、旅者981544、ChenweiSmile和kdd的名片，附上他们所擅长的领域等信息，以及推送“C++程序设计基础”课程中关于“函数”和“递归”的相关章节的资料。

图 4-2 聚类分析结果

至此，再次梳理本文所介绍的基于SLN的数据挖掘过程：

第一步，通过基于Python 2.7的Scrapy 3.0框架的网络爬虫从目标SLN的论坛上获取足量的数据。

第二步，通过数据预处理过程进行清洗、提纯数据，然后通过分词程序分析数据，提取关键词并建立用户词典。

第三步，通过对文本数据的关联分析，提取出关键词信息之间的关联度。

第四步，以用户和关键词为节点，以节点之间的关联程度为权值建立SLN图论模型。

第五步，通过对SLN图论模型的聚类分析，把图中的节点合理地划分为若干簇。

我们在SLN上杂乱无章的论坛挖掘出了数据，并按相互关联程度的紧密或疏远把所有的用户和信息分成了若干大小不一的群体。当一个新用户加入SLN时，无论他是什么方式进入到一个领域或课程时，我们都能根据上述数据挖掘成果，把用户最需要的信息推荐给他。

# 第五章 总结

本文介绍了对在线社会学系统进行大数据挖掘的一种方案，通过收集大量或全部数据，对其进行分析、演化、建模和挖掘，提高向新用户推送信息的效率。文中介绍了参与到这个过程中需要用到的部分知识或工具，包括基础图论及相关算法的实现、数据挖掘和机器学习的基础知识、关联分析和聚类分析的基础知识和相关算法的实现、网络爬虫技术、文本处理技术、分词技术等。

# 参考文献

[1] Pan-Ning Tan, Michael Steinbach, Vipin Kumar. 范明，范宏建等译. 数据挖掘导论：完整版（第2版）. 人民邮电出版社. 2011：201-216, 306-315

[2] Brinton Christopher, Chiang Mung. Social Learning Networks: A Brief Survey. 2014 48th Annual Conference on. IEEE, 2014: 587-592

[3] WW Moe, PS Fader. Capturing Evolving Visit Behavior in Clickstream Data. Journal of Interactive Marketing18(1). 2004: 5-19

[4] Ting I-Hsien, Kimble Chris, Kudenko Daniel. UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a Web Site's Design. IEEE/WIC/ACM International Conference on Web Intelligence. 2005, 9: 179–185

[5] Blodget Henry. Compete CEO: ISPs Sell Clickstreams For $5 A Month. Seeking Alpha. 2007, 3

[6] Chatterjee Patrali, Hoffman Donna, Novak Thomas. [Modeling the Clickstream: Implications for Web-Based Advertising Efforts](http://www.jstor.org/pss/4129736). Marketing Science.22(4). 2003,Autumn. 520-541

[7] Sweiger Mark, Madsen Mark R., Langston Jimmy等. 陆昌辉，张光剑，陈佐等译. 点击流数据仓库. 电子工业出版社. 2004年1月: 28-64，202-210

[8] Romero Cristóbal. Ventura Sebastián. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews. 40(6).2010: 601-618

[9] 贾振声，徐文权. 树形图求中心点及其算法. 太原科技大学学报. 30（5）.2009，10:405-408

[10] Cormen Thomas, Leiserson Charles, Rivest Ronald 等. 潘金贵，顾铁成，李成法等译. 算法导论.原书第二版. 机械工业出版社. 2006，9：322-335, 365-369

[11] 高润霞. 图的直径与最小特征值[学位论文]. 安徽大学, 2010.

[12] 丁振国. 陈静. 基于关联规则的个性化推荐系统. 计算机集成制造系统CIMS. 9(10). 2003, 10: 891-893

[13] T Rolleke, J Wang. TF-IDF Uncovered: A Study of Theories and Probabilitties. In. International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore. 2008: 435-442

[14] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts. In. Association for Computational Linguistics. Barcelona, Spain. 2004: 404-411

# 致 谢

感谢纪晓东老师对我的毕业设计提供的帮助和指导。在毕业设计期间，纪晓东老师多次对我的论文提出宝贵的意见和建议，并一直督促我认真完成论文所需的工作，去粗取精，完善论文结构和内容。

感谢彭木根老师、赵中原老师对我的毕业设计的指导和关心，尤其是在论文框架和进度安排上提出了重要的建议。

感谢孔晓旻同学在毕业设计期间无私地给我提供了大量的知识信息和参考资料，以及课题研究所需的工具、技术和研究方法。

感谢本文所涉及全部知识、信息、参考文献的作者、研究人员和学者，正是他们的研究成果使本文获益匪浅。

感谢本文所用到的所有开源平台和数据处理工具的项目发起者、开发者和维护者，使本文在数据解析、处理和可视化上节省了大量的时间。

由于作者水平有限，本文难免有不足之处，恳请各位老师、同学和读者批评指正！