

**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**

在线社会学习系统是一种包含网络教学行为的社交网络。在线社会学习分为：视频点播、直播、慕课和反转课堂等。目前，在线社会学习系统快速发展，成为当下和未来的一种重要教育模式。随着教学模式不断丰富，学生和教师参与度不断增加，教学效果也越来越好。

在线社会学习系统既保留了传统教学行为过程中的许多特点，又包含很多社交网络的属性。比如在教学过程中包括四种学习模式：课堂学习、课文，考试以及答疑，对应在线系统上的视频学习、课文资源、测评以及论坛；而在课程的选择、课堂表现、参与论坛等方面又具有显著的社交网络特征。用户在系统上产生的数据大量地体现出这两方面的特点。

本文通过对在线社会学习系统的数据进行分析和挖掘来研究用户使用系统的过程中与知识和信息之间建立起来的内在和外在的、具体和抽象的联系，并将其转化为图论模型。通过对这个模型的分析，系统能够高效地为用户呈现所需信息。

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**

Social Learning Network (SLN) is a kind of social networking system contains online lectures. There are several types of SLN, which are lecture videos, lives, MOOCs (Massive Open Online Courses) and FLIPs, etc. So far, SLNs have been developing drastically and have become an important module of education. As the learning modules are diversified, students and teachers participate more and the experiences are enhanced a lot.

SLNs have reserved many features of traditional teaching behavior; meanwhile, they have derived many attributes from Social Networking Systems. For example, there are usually 4 modalities of traditional classes which are lectures, textbooks, exams and Q&A corresponds to videos, text resources, quizzes and forums in SLNs. As for course choosing, class behaviors and forum behaviors which apparently reflect the signature of social networks. Massive user data on SLNs shows these two aspects of systems.

This article shows the researches on how internal/external, abstract/specific connections had been established during the time when the Social Learning Systems were using by analyzing and Data Mining on them, and turn the connections into a Graphic Model. Thus, through the analysis of this model the system to allowed present the information that users needed.

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

**目 录**

\*\*\*\* 待添加 \*\*\*\*

1. **绪论**
   1. **课题简介**
      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，一般说来是一部分人通过在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，教学者和学习者之间、教学者和学习者内部都不是隔绝的，相反，随着时间的推移，这些角色之间将会反复地互相获得知识、传播知识，甚至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。

SLN所保有的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具有极强的线下沉淀，许多学校和教师利用SLN完成部分教学任务，甚至用SLN作为家庭作业来完成主要教学任务，而课堂上的时间用来答疑。

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两种主要的形式：MOOC（MOOC， Massive Open Online Courses大规模开放在线课程）和FLIP（翻转课堂）。

MOOC，又称慕课，是一种SLN的平台化形式，突出规模化、开放性、不受时空限制等几大特点。目前，国内外有十几个具有很大影响力的MOOC平台，如“慕课三巨头” Coursera，edX，Udacity；国内的各大学MOOC（如清华大学研发的学堂在线），斯坦福在线(Stanford Online)，Open2Study（澳洲最大MOOC平台），Ewant（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的MOOC平台），慕课网（imooc），酷学习（kuxuexi）等。这些应用坚持将大量免费的高质量内容开放给所有用户，不仅让有条件的用户获得优质的知识服务，甚至帮助了很多落后偏远、缺乏师资的地区的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，对于有相关方面需求的学校、企业提供了免费的远程培训的标准流程。

MOOC平台作为获得教育资源的一种方式，一般仅作为课余或业余的补充。

除了平台化的MOOC外，还有一种SLN的重要形式——翻转课堂（FLIP，又称Flipped Classroom或Inverted Classroom）,指重新分配教学时间的一种模式。在这种模式下，宝贵的课堂时间给予学生来专注于个性化、定制化、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、听播客、在网络上与其他同学讨论等。翻转课堂是一种混合式的教育方式，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。在互联网时代，学生通过SLN能够获得丰富的在线课程资源，包括优质的视频、播客、图文并茂的有趣电子书（相比枯燥的纸质课本来说）等。

翻转课堂作为国民教育序列的一种方式，将对学生和老师、内容和模式都提出较高的要求。在学生自主性较差、老师教学灵活性偏低、内容制作和模式设计的水平都不高的情况下，翻转课堂的教学效果很难得到基本保证。总体来说，翻转课堂模式在中国发展情况不如MOOC平台。

* + 1. **研究内容**

随着SLN的发展日新月异，对于更好用的功能、更优质的内容、更高效的学习模式的需求越来越强烈。来自不同国家和地区、不同年龄段、拥有不同需求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，规模大幅度增加，有些热门课程甚至有全球数以十万计的用户在同时观看。对于这些庞大的数据，已有许多研究成果。

目前关于SLN的研究主要集中在以下几方面:（一）发展沿革，研究SLN的起源、发展和现状，通过调查、试用、数据挖掘等研究方式探究SLN未来的发展方向；(二)激励体制，探究用户对于SLN的需求，主要研究MOOC平台和FLIP如何鼓励用户参与学习和延长用户的使用周期；（三）用户体验，通过分析点击流数据来帮助改进网站使用方法，增强用户体验；（四）网络模型，通过分析用户资料和论坛数据等来构建SLN的网络模型，研究用户和知识、信息之间的互动关系。

* 1. **用户数据简介**
     1. **社交网络模型**

SLN的网络模型可用图论模型来构造。图中的节点（node）代表学习者或课程，图中的边（link）代表这些学习者和课程相互之间的联系。这些边可以被赋权，用来标识它所代表的含义的重要程度。这些图论要素都可以从实际的SLN上抽取出来。一般来说，SLN的图论模型分为以下四种：

**学习者之间的无向图**

如果把学习者单独拿出来作为图的节点，我们可以用无向图的边来表示两个学习者之间是否有一些相同的特质。这些特质可以是年龄、地理位置、教育背景或是否选择了某个（些）课程等。对于节点和来说，有[2]：



公式 1-1

说明：如果和同时满足包含特质，那么，反之为0；为全部特质的集合。这个公式的意义是:若两个学习者之间同时满足至少个特质，那么在SLN的图论模型中这两个节点之间就会有一条无向边。

**学习者之间的有向图**

在研究信息流的过程中需要用到这样一种模型。我们从论坛的互动中可以提取出用户之间的关系。如果用户回了用户发的帖子，那么就会有一条从指向的的单向边。这样的规则可以进一步严格为只有“最佳回帖”才能生成一条单向边。依据这种规则生成的很有可能是一个多重图（Multigraph），因为某个用户可能和另一个用户之间产生多次符合条件的交互。

**包含学习者和关键词的无向图**

信息和知识本身当然也可以成为SLN图论模型中的另一种节点。能够成为节点的关键词或关键信息能够以多种方式提取出来，如利用提取关键词的相关工具对讨论区中的文本数据进行分析，或利用教学者提供的教学大纲来提取。定义在学习者节点之间的边与上文第一种无异，定义在学习者节点和关键词节点之间的边定义为该学习者是否在讨论区中回复了足够次数的关于该关键词的帖子。

**包含学习者和关键词的有向图**

这种图揭示了通过发帖和跟帖，知识和信息在用户之间是如何流动的，而且这种图更可以分割为学习者-关键词的二部图。具体规则为：针对每一个帖子，生成一条从发帖人节点指向该帖关键词节点的单向边，之后每一个回帖生成从该帖关键词节点指向回帖人节点的单向边；另外，每个回帖中生成一条由回帖人指向回帖关键词的单向边。这些单向边的权重可由该边所代表的发帖或回帖的赞成/反对数来换算。

以上四个例子是定义在SLN上的一些很实用的分析模型，配合图论的相关算法可以起到很好的数据挖掘作用。

* + 1. **点击流数据**

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。用户在页面上点击一下鼠标后，这个动作将会被记录在客户端或服务器上，也有可能会被记录在浏览器、路由器、代理服务器上。点击流分析是一种非常实用的用户行为分析方法，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。

原始的点击流数据可以从服务器上收集，点击流是一系列页面请求，每一个页面请求产生一个信号。经过一些特殊的鉴别算法，可以筛选出那些真正由人类用户的访问和使用产生的信号。通过跟踪这些点击流信号，网站管理员能够监测到用户的使用行为[4]。点击流数据主要应用于广义的市场调查，通过分析点击流，任何拥有网站的管理员、研究人员、博主和个人都能得知如何改进他们的网站。

使用点击流数据会涉及到隐私权问题，一些公司会出资够买用户的点击流数据，价格大约为每个用户每个月40美分[5]。未经授权的点击流收集被视为后门软件或间谍行为，如要使用授权的点击流数据，也要通过有该项权限的公司为那些愿意分享点击流数据的用户安装收集点击流数据的程序来获得。因此，点击流数据很难获得，不适合作为第三方分析的目标，也不是本文所介绍工作的重点。

* + 1. **论坛数据**

MOOC平台上的讨论区（论坛 ）形式结构上等同于一个问答网站，如各大贴吧、BBS和知乎。一般来说，一个讨论区包含若干发帖，每个发帖后有若干跟帖，每个发帖和跟帖可以赞同（或称“顶”）或反对（或称“踩”），形成了一个三级结构。论坛和讨论区的用户可以发新帖，跟帖、赞同或反对已有的帖子。每个课程后面的帖子一般有几种类别和许多标签，但是很多时候发帖的内容和分类、标签内容无关或根本就没有分类和标签。有的系统上也并不存在赞同和反对的功能。

在MOOC上，学生交流的话题一般不仅限于对学习内容的讨论，而且容易转化为闲谈。在翻转课堂的模式下，学生用户在SLN上的交谈讨论不像在MOOC上那么典型，一方面是因为学生之间以及学生和老师之间在线下的交流已经满足了这方面的需求，另一方面是因为线上的监管人员会限制这样的行为。

在本文所介绍的数据挖掘工作中，主要针对论坛内发帖、跟帖的主体文字内容，不关心分类、标签（相关度低）以及赞同数和反对数（不具有普遍性）。这些文字内容进行文本数据挖掘后，能够清晰地看出用户所关注内容和反馈的焦点，是用户最直接的使用体验，也是系统上对于内容最精确的数据。这些文本数据挖掘的结果既可用于描述性任务，也可用于预测性任务，是目前SLN研究工作的重点之一。

* 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，也无法用传统工具和技术来获取有用信息。此时就需要开发新的、互联网式的数据分析方法，即数据挖掘技术。

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。

目前，在商务、医学、科学和工程方面均有不同层次和规模的数据挖掘应用。如：小到针对消费者购买数据合理安排商品在货架上的摆放位置或推送其他商品的广告；大到利用地球轨道卫星收集和探测到的海洋、陆地、大气等庞大的复杂数据来推算生态环境等问题。这些任务的共同点是:在大型数据库中自动地发现有用的信息。这个过程被称为数据库中知识发现（Knowledge Discovery in Database）。

如图1-1，输入数据可以是各种形式和类型如纯文本、电子表格或数据库等，并且可以实现分布式存储。输入的数据经过数据预处理（Preprocessing）过程将原始数据转换和分类为可以被分析的形式和类型。这个过程可能包含清洗数据，即去除噪声等无效数据、去除重复数据、去除与当前任务无关的数据等。这个过程所耗费的时间和脑力成本可能是整个KDD过程中最多的。

将数据去粗取精后，进行数据挖掘过程。数据挖掘的结果将会被送入决策支持系统中，例如企业的商务智能（Business Intelligence，BI）系统，这个过程称为“结束循环”（Closing the loop）。这个过程中，数据挖掘结果需要进行“后处理”（Postprocessing），将有用的和有效的结果留下来。如“数据可视化”将这些结果转化为各种图表，使决策人员能从各种不同视角分析判断。

传统数据分析技术针对新型海量数据主要遇到可伸缩性、高维性、异种性与复杂性、数据所有权与分布、非传统性等问题。为解决这些问题，研究者们依托已有的数据分析领域的相关方法和研究结果，开发了一系列数据挖掘手段。其中，数据挖掘技术主要利用了统计学、模式识别、机器学习、搜索算法和建模技术的相关理论。

图 1-1[1] KDD过程示意图

数据预处理

数据挖掘

后处理

信息

输入数据

特征选择

维归约

规范化

选择数据子集

选择数据子集

模式过滤

可视化

模式表示

可视化

模式表示

选择数据子集

数据挖掘的两大主要任务是:预测任务，即根据已有数据预测特定属性的数据；描述任务，即找出数据中的潜在联系，通常需要后处理来验证和解释。这两大任务又分为四类具体的任务，即聚类分析、关联分析、预测建模、异常检测。在SLN的数据挖掘任务中，主要通过聚类分析和关联分析来获取数据库中的有用信息(描述任务)，然后通过预测建模来进行推测某个用户接下来的行为（预测任务）。

1. **在线社会学习系统数据分析**

**2.1 论坛文本数据分析**

针对论坛数据的分析，其目的和用途有很多种，其中主要的如对论坛文本数据进行分析以提取关键信息，或对针对论坛的社交情况建立合适的SLN图论模型。这一节仅讨论对论坛纯文本数据的理论分析，本章的下一节将重点讨论由论坛数据提取社交网络模型的相关方法。

正如前文所介绍，SLN的论坛普遍是一种典型的、简单的、附在主体课程后的“讨论区”，它拥有一个鲜明的三级结构(如图2-1所示)，每一个发帖、跟帖都可以看作是一次针对某个关键信息的独立发言。简单地，一个发言可以仅有作者和内容；然而大部分讨论区或论坛一般都会给发帖者和跟帖者以及“仅仅路过”的用户一些独特的权力。

图 2-1 讨论区结构示意图

课程名称：XXXX

讨论区

发帖人：用户1 帖子分类：分类1，分类2，…… 123赞成 456反对

\*\*发帖内容\*\*

标签：标签1，标签2，…… 回复

跟帖作者：用户2 1赞成2反对

\*\*发帖内容\*\*

标签：标签3，标签4，…… 回复

跟帖作者：用户3 3赞成4反对

引用 用户2 的回复：

\*\*引用内容\*\*

-------------------------------------------------------

\*\*发帖内容\*\*

标签：标签4，标签5，…… 回复

……

比如，发帖者可以给自己的帖子内容分类或贴标签，以保证回帖的内容在期望的范围内；有些论坛允许发帖者删除自己帖子下的回帖内容，以明确这项权力，通过用户自觉自愿的方式来维护论坛内容的质量。很显然，当一个SLN用户真正想要寻求帮助或发布信息时，他是肯定不会允许别人在他的地盘上捣乱的。而论坛管理员或该讨论区所属课程的老师、助教等管理角色的工作之一就是保证这些发帖人不是来捣乱的。

回帖最主要的功能亮点是可以引述前文，包括发帖和之前回帖的内容，一篇帖子被引述的次数越多，说明他的观点越具有影响力，不论是正面的还是反面的。与发帖相同，回帖的内容也可以被贴上一些标签，用来更好地被检索、发表看法或指出关键信息等。当然，如果没有这项设定也不大影响他们的回复。

最后，所有用户都可以对已有的内容，包括发帖和回帖进行投票，可以投赞成票或反对票。投票、分类、标签甚至作者对内容的分析结果并不构成关键的影响。

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。

如图2-2，数据预处理的主要流程是：首先，通过数据采集手段从网站上收集相关的论坛信息；其次，将混杂格式的数据（除了纯文本格式外，还有XML、JSON等格式，一般包含GBK或UTF-8编码格式，涉及到编码转换问题）变为统一形式、统一编码格式的纯文本数据；第三，将这些纯文本字符串中的无效、无用、无价值字符以及被我们以先验方式屏蔽的数据去掉，如英文字母、符号、公式、图片、错误信息（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），将专有名词加入到用户字典中，比如“柯西不等式”等；最后，将所得数据送入分词、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。

图 2-2 数据预处理过程

输入数据

**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

文本数据挖掘过程主要进行关键词的关联分析和聚类分析，通过这些分析手段能够获得关键信息之间潜在的联系。这一部分将在第四章进行详细介绍。

**2.2 社交网络模型**

用户在SLN上的互动是一种典型的图论模型。在这个模型中，用户和知识、信息都可以作为结点，节点之间的边（可以是有向边或无向边），代表这些用户、知识之间的联系。上文介绍过，SLN社交网络模型有4种，即：

·用户之间的无向图，用于提取用户之间的相似性 (I)

·用户之间的有向图，用户提取用户之间的互动性 (II)

·用户与信息之间的无向图，用于提取用户对包含该信息话题的参与度（III）

·用户与信息之间的有向图，用于提取用户对包含该信息话题的贡献度（IV）

把这四种模型中节点之间的联系方式的精髓结合起来，可以形成一个混合图，既包含有向边又包含无向边。无向边表示节点之间一种平等的相互关系，比如模型I中，用户之间的相似度，即用户的静态资料中显示的共同点的个数大于*P*时，有这样一条无向边。有向边表示一种动作，如用户指向信息的边，表示该用户发表过包含该信息关键词的帖子；信息指向用户的边，表示该用户回复过多少包含该信息关键词的帖子。把这两种计量放在一起，需要调整一下各自的权重,使其能用一个标准衡量。如果在某个建模任务中，用户之间的相似度更为重要，那么可以增加用户之间每个共同点的权值；反之，增加用户每次发帖和回帖所占的权值即可。

此外，投票和跟帖数和评论数可以帮助我们更好地了解这篇帖子在讨论区中所占的权重，以及在该帖子所包含的关键词集合覆盖的SLN网络模型中所占的权重。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，1000人反对的帖子，他的影响力和价值远远超过一篇有100人赞成，50人反对的帖子。

由图论基本知识可知，无向边可以转换为双向边，则可以把这个混合图变为一个有向图。反之亦然，可以把有向图转化为无向图，仅研究节点之间的相关性，不研究拓扑次序。在这个有向图的基础上，可以利用许多图论算法对其进行分析。如最短路算法、合并强连通分量、查找直径、查找中心点等，在第三章中将会详细介绍。

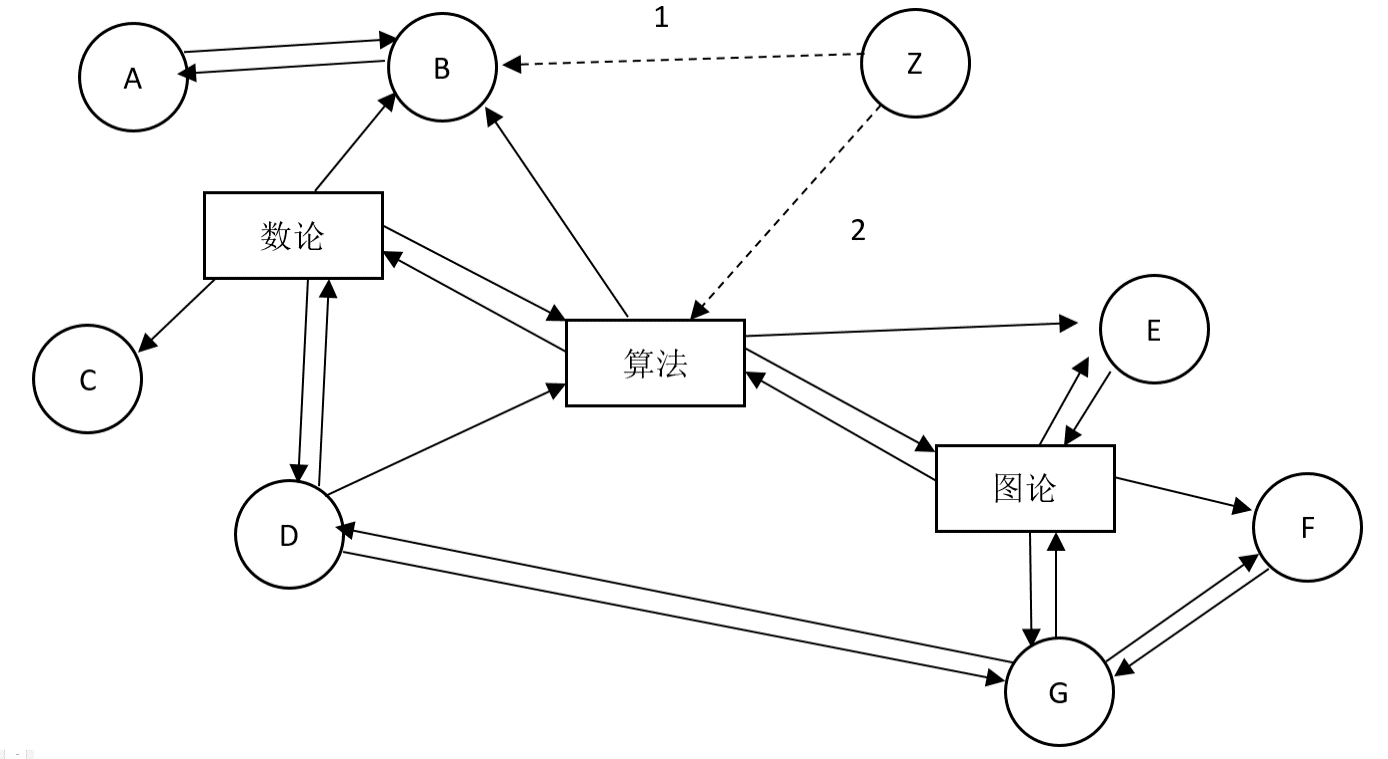
如图2-3是从某SLN的某课程讨论区中提取出来的网络模型。

图 2-3

其中，圆圈内含大写字母表示用户节点，矩形框内含中文单词表示关键词。

由图可知，用户D发表关于数论的帖子，得到了B、C、D的回帖；D发表了关于算法的帖子，得到了B和E的回帖；E、G发表了关于图论的帖子，得到了E、F、G的回帖。从先验的角度得到，算法、数论和图论之间具有千丝万缕的关系。A与B、D与G之间满足用户之间的关联条件。

注意一点:在生成子图时，某一条边被选中当且仅当它的权值大于等于一个权重阈值。

现在使用强连通分量合并算法处理该子图（又称“缩顶点”，即把一个极大强连通分量缩为一个超级顶点，去掉分量内部的边，保留跨分量的边），得到如图2-4的子图。

现有一个新加入该子图的用户Z（与图中其他节点没有任何关联），考虑如下两种情况:

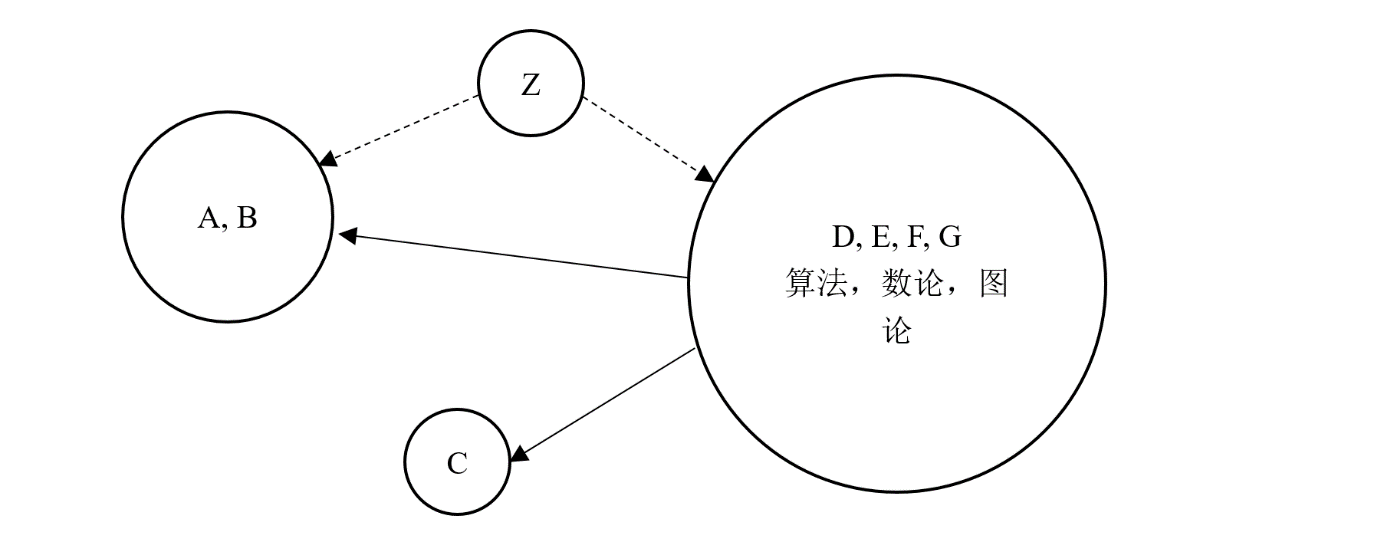
1. 用户Z通过与用户B互动加入该子图，如互相关注、发私信或通过该用户的邀请注册等。

图 2-4 缩顶点后的子图

在这种情况下，用户Z在该子图中所包含的关键词中没有明确的学习目标，而拥有许多“志同道合”的其他用户（如用户B），这是一个带有先验性质的条件。在这种情况下，应该向用户Z推送它有可能感兴趣的课程。在该子图中，“数论”的度为6，“算法”的度为7，“图论”的度为7；而“算法”距离用户B的距离比“图论”更近，所以推荐时，应最先推荐“算法”。

1. 用户Z通过浏览包含“算法”关键词的网页或课程加入该子图。

这种情况下，用户Z一般带着比较明确的学习目标，此时再向其推荐其他课程，效率可能会降低。就算用户Z选择加入我们推送的课程，也未免不会影响他的完成率。如果向其推荐一些志同道合的其他用户，可以增强用户在SLN上的互动性。

如在该子图中，我们可以向其推荐该强连通分量中最活跃的用户。在不考虑权重的情况下，用户G的度为6，比其他用户的度都高，所以应最先推荐用户G的名片。

**2.3 点击流数据分析**

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，而这些数据所涵盖的范围远远小于我们所需要的基于用户行为的点击流数据。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。

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，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）。

在SLN中，这种用途可以具体化为：根据用户的资料推荐他可能感兴趣的课程或其他用户，以及根据用户已购买或参与（完成）的课程向其推荐相关的课程。这相当于数据挖掘中的“预测任务”。与此同时，这其中的“描述任务”可具体化为：什么样的课程最受欢迎？什么样的人群最想要学习哪些课程？某个课程中学生会在哪个时间退出观看？这些问题都是可以通过点击流分析和数据挖掘技术得到答案的，也是网站研发人员和课程设计人员能够精益求精的重要依据。

建立一个点击流数据仓库的第一步，就是在网站服务器所产生的日志文件中挑选针对用户点击的各种维度的数据组成一条条点击流数据，然后还原网站用户的行为，并最终得到导致这些行为的原因。其次，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。最后，经过清洗和转换的点击流数据存储在点击流数据仓库中。因此，点击流数据仓库中含有大量历史数据，能够支持许多与时间有关的分析任务。

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。

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；对于第三方来说，除开分析和获取数据的产权成本，真正高质量、有价值的数据在全部数据中的占比非常低，而且有可能在某种程度上增加数据的误差，分析起来存在一定难度。

1. **数据挖掘技术**

数据挖掘技术脱胎于传统数据分析方法，突破了传统数据分析方法无法分析非传统类型数据或海量数据的局限性。在利用数据挖掘技术获得信息之前，首先应通过一定技术手段获取足量的数据；其次，应对收集到的数据进行预处理，使其符合可挖掘的格式和类型。本章第一节介绍利用网络爬虫收集数据的具体方法；第二节介绍数据预处理方法；第三、四节介绍数据挖掘技术中的关键技术：关联分析和聚类分析。

通过数据挖掘技术产生的数据元素之间的潜在关联将作为构建SLN图论模型的重要数据来源。

**3.1 数据收集**

对于网站管理员、数据分析员、商务智能人员等角色来说，获取数据的方法比较直接，大部分用户数据可直接从数据库获取，这部分数据的类型、格式、编码等都比较符合数据挖掘的要求。对于实时数据，需要利用站长工具等方式监控网络状态等测度。对于第三方分析来说，后台数据和服务器日志无法获取，可以通过网络爬虫自动获取网页上的数据。

网络爬虫是自动提取网页内容的程序。在本文所介绍的工作中，主要分析挖掘SLN的论坛数据，选用基于Python2.7 Scrapy3网络爬虫框架。程序清单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。

**程序清单3-1 网络爬虫算法**

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

网络爬虫是搜索引擎的重要组成部分。大型的并行分布式网络爬虫将会有巨大的资源开销，也会占用很大的带宽，影响被爬取网站服务器的性能。在实际操作的过程中，绝大部分网站包括SLN都需要用户先登录才能访问，或者用识别同一IP、同一浏览器、访问间隔时间等方式来鉴别网络爬虫，从而拒绝访问。对于登录验证来说，可以使用在程序的前段加入Cookies的方法来自动通过验证；对于识别统一IP或浏览器来说，应建立相应的IP池，随机更换每次请求的浏览器；最后，应设置每次的访问间隔在一定范围内随机，如0.5s~1.5s，既要保证随机的幅度避免被服务器识别、屏蔽，又要兼顾程序执行效率。

表3-1给出了本文所介绍的SLN大数据分析和网络建模相关工作所需的字段。

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、编码都不一定统一，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一些预处理，分析出网络中的关键信息，最后才能送入建模过程。

|  |  |  |  |
| --- | --- | --- | --- |
| 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（包括但不仅限于）

**3.2 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2-2的流程，进行数据清洗、数据提纯以及文本分析的程序。

**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），而当下中文WEB站点基本使用UTF-8格式，所以最好编码为UTF-8格式。

**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对文本数据分析程序毫无意义的字符。在中文MOOC分析中，假定所有ASCII码表内的字符，包括但不仅限于大小写英文字母，数字，所有英文标点符号、表情符号，以及所有中文标点符号。中文数字按此规则也应去除，但由于在分词程序中，中文数字一般不被视为单词，且包含同一数字的词汇重复率极低，所以在不考虑系统开销的情况下，可以暂时不用担心。

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（甚至比某些关键信息出现次数还多），对文本分析来说无意义或者有害。如各类语气词、象声词、虚词等。

去除停用词的高质量数据称为有效数据。

**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，它完整地保留了原始数据的语义信息，尽量剔除了垃圾数据。文本分析程序有两个功能，第一是分词，把连续的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。第二是提取关键词，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3-1给出了文本预处理每一步的效果，数据来自国内某著名MOOC平台，数据经过归一化处理。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。

图 3-1 数据预处理步骤

数据源

∞

原始数据

5.52

高质量数据

4.92

有效数据

3.19

用户词典

1

网络爬虫

去除无效字符

去除错误信息

用户词典 0.044

停用词典 0.343

分词

提取关键词

**3.3 关联分析**

在之前的文本分析程序中，我们把原始数据分解为一系列单词。前文介绍的图论模型中，用户和关键词之间具有一些具体联系，如图3-2所示:

关键词1

关键词2

关键词3

回帖关键词1

回帖关键词2

回帖关键词3

图 3-2图论模型示意图

其中，虚线双向箭头表示抽象联系，如表示利用公式1-1所确定的用户之间的关联，表示通过文本数据挖掘过程所确定的关键词之间的关联。实线单箭头表示具体联系，如表示发帖作者指向发帖关键词的单向边，发帖关键词指向回帖作者的单向边。

在本节和下节中，主要任务是找寻关键词之间的抽象关系，即所表示的关系。在之前的工作中，我们把每一个关键词抽象为一个“项”，每一个帖子抽象为一个“事件”，包含这个帖子内容中的所有关键词的“项”。“项集”是一个组合概念，包含若干个“项”的集合。

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。

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3-1** 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]。

**程序清单3-2 *Apriori*算法产生繁项集**

1: size=1 #项集大小

2: {Fsize}={ {i} | i{I} & sup(i)minsup } #sup()为项或项集的支持度

3: REPEAT

4: size=size+1

5: FOR {C} : {Fsize-1} #扫描每个项集

6: FOR x : {I} #扫描每个项

7: IF sup({Ci})minsup #支持度条件成立

8: {Ci}{Fsize} #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

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。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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图表3-1所示。

图表 3-1 项集数规模走势

由图示可以看出，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3-2** 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3-2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 3-2 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。

**例3-3**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3-3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 3-3 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”。

从以上两个例子可以看出，在SLN图论建模中，项(或项集)的支持度即为节点自身的权重；规则置信度即为关键词与关键词之间建立边的权值。

**3.4 聚类分析**

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。
2. 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。
3. 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。

另外，我们所实现的聚类分析与机器学习的聚类分析有所不同，原因是（1）在本文所介绍的SLN的关键词数据挖掘过程中，不存在训练的过程；（2）SLN模型是一个图论模型，无法直接套用普通的聚类分析过程。

下面介绍一种简单的聚类方法——K均值算法。K均值算法比较简单，首先，选择K个初始质心，每个点选择最近的质心，选择了同一个质心的点标记为一个簇。然后，根据每个簇的点的信息，更新每个簇的质心。重复这个步骤，直到质心不发生变化为止（或到仅有很少一部分点可能会改变质心为止，如1%的点）。

把K均值算法与SLN图论模型结合起来，就能很好地完成聚类任务。影响K均值算法效果和效率的关键因素有四个：（1）如何确定K值及初始质心?（2）如何快速计算项到质心的距离？（3）如何根据点集快速重新选择质心？一个可行的解决方案如下：

**K值和初始质心的确定**

K均值算法的核心就是K的值，它代表了聚类分析的强度。如果K不合适，则会产生错误聚类，即把本该属于一个簇的项分隔开，或把不属于一个簇的项强行聚类在一起。另外，为了是质心有意义，规定质心必须从图论模型中的节点集内选取。

在聚类分析开始前，我们的预期中，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，分为两个簇。

根据这个设想和图论模型的性质，我们可以设计一个染色算法：设置一个阈值，扫描图中的每个节点，假如某节点还没有被分类，则从这个节点开始遍历。遍历时，如果某节点所连的某条边的权值大于我们设置的阈值(换言之，即两个节点之间的距离小于对应的阈值)，那么将这条边的另一端连接的节点加入待遍历的点集，并把它与遍历的起点分为同一个簇。把所有的节点都标记完后，标记的簇数即为K。

在这个算法中，对于一个基本静止的图（不会突然有巨大变动的图，例如在两次数据挖掘之间有99%以上的节点和边的状态保持不变的图）来说，一个阈值对应一个聚类结果。如果某个阈值对应的聚类结果中，发生了上述错误聚类，那么就调整K值重新执行染色算法。在一个相对合理的染色结果和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。

初始质心没有最终的意义，在我们用染色算法分号初始的簇后，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。

**计算项到质心的距离**

确定好质心后，如何计算对于每个项最近的质心呢？不妨把问题转化一下，计算每个质心到图中每个点的最短距离，在此过程中，每个点维护目前离自己最近的质心的编号和距离值，如果当前质心到该点的距离小于其维护值，则更新这个维护值。单源最短路问题使用Dijkstra算法，或更高级的SPFA、A\*算法均可。

该算法过程的伪代码在程序清单3-3中给出：

**程序清单3-3 renew过程：计算质心到所有节点的距离并确定节点的簇**

1: FOR i: 1 TO N

2: Hi=null # 初始化每个节点所属的质心

3: Di=∞ # 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)→{Disi | 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 Di>Disi THEN # 更新维护值

9: Di←Disi

10:Hi←heart

11: END IF

12: END FOR

13: END FOR

**根据点集重新选择质心**

根据点击重新选择质心的过程，就是在点集所对应的连通子图中找到中心点。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。

引入一个概念：图的直径。在一个图中，图的直径是指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(有时指最短路径的长度，即最短距离)，直径可能有多条。为解决这个问题，不加证明地引入两个定理：

**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。

**定理2**连通图的质心一定在图的某一条直径上，具体是哪条直径没有区别

通过这两个定理，这个问题就转化成了：在点集所代表的子图中找出一条直径，在这条直径上找出一个点，该点与图中其他点的距离和最小，则该节点为所属子图的中心点，即所属点集（簇）的质心。具体算法过程如下：

通过定理1，我们能够设计一个算法，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。

通过定理2，我们能够设计一个算法，在从直径的一端向另一端逐序遍历每个节点，计算该节点的距离和，找出使距离和最小的节点即可。特别的，如果以直径的一端到另一端的节点顺序（离散值）为横轴，以距离和的值为离散函数值的话，将形成一个下凸函数。所以随着这个过程的运行，距离和的值会先减小，后增大。一旦其开始增大，就不会再减小了。所以增大前的最后一个值即为该离散函数的最小值，也就是我们所求的结果。

K均值的伪代码如程序清单3-4所示。

**程序清单3-4 K均值算法**

1: {Heart}←dye {I} #利用初始化染色算法确定K和初始簇、初始质心

2: REPEAT

3: renew # 调用renew过程更新所有点所属的簇

4: FOR {Ii} : {I} # 遍历每一个初始簇

5: {Di}←diameter {Ii} # 通过上述两次最短路算法获得直径的顺序路径

6: FOR d : {Di} # 按顺序遍历直径上的每个点

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

8: mid\_sod←sum\_of\_distance ( d→{ Ii} )

9: Hearti←d # 更新质心集

10: END IF

11: END FOR

12: END FOR

13: UNTIL quit\_condition=TRUE # 直到上述停止条件为真

通过我们“定制”的算法可以初步将SLN图论模型划分为K个聚类。

至此，再次梳理本章所介绍的基于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，以用户和第二步建立的词典为节点，以用户相关度公式、用户发表包含某关键词的帖子的频度以及第三步提取的关联度为边的权值建立SLN图论模型，并通过归一化使权值符合同一个衡量标准。

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。这意味着，经过这五步过程，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，无论他是通过另一个用户、还是通过某个页面链接进入到一个陌生领域或课程时，我们都能根据上述数据挖掘成果，把该领域内最应首先学习的课程和最资深的用户推荐给他。

在下一章中将具体对某MOOC平台的部分课程进行上述数据挖掘过程并展示其结果。

1. **数据挖掘成果**

\*\*\*\*待添加\*\*\*\*

1. **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，即通过收集一定量（少量、大量或全部均可，视任务需求和开销而定）的数据，对其进行数据分析、演化、建模和挖掘，以达到能够科学高效地向新用户推送其他用户和课程的目的。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。

附录：参考文献(未格式化！！！！)

[1] 数据挖掘导论

[2] Brinton, Christopher G., and Mung Chiang. "Social learning networks: A brief survey." Information Sciences and Systems (CISS), 2014 48th Annual Conference on. IEEE, 2014.

[3] WW Moe, PS Fader (2004),“Capturing Evolving Visit Behavior in Clickstream Data” Journal of Interactive Marketing (2004)

[4] Ting, I. H.; Kimble C; Kudenko. D (September 2005). "UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a Web Site's Design." (PDF). IEEE/WIC/ACM International Conference on Web Intelligence: 179–185. doi:10.1109/WI.2005.153.

[5] "Compete CEO: ISPs Sell Clickstreams For $5 A Month". Seeking Alpha. 2007-03-13. Retrieved 2011-09-15.

[6] Educational Data Mining: A Review of the State-of-the-Art

[7] 点击流数据仓库

[8] Educational Data Mining: A Review of the State-of-the-Art

[9] Learning Analytics: Definitions, Processes and Potential

[10] http://www.docin.com/p-1062977464.html 这篇文章有我要的图论证明！