基于高校图书馆的学生借阅行为分析报告

学校名称 广东外语外贸大学

报告成员 张礼明、林家骏、赵彦青、孙博轩

学科专业 软件工程

起止日期 2015 年 11 月 16 日--2015 年 12 月 20 日

提交日期 2015 年 12 月 21 日

摘要

随着计算机技术的迅速发展和网络技术的广泛应用，电子信息化的现象渗透到各个行业与领域之中。图书馆的借阅管理已经由传统的人工记账方式转化为计算机信息系统智能管理。而图书馆系统在长期运作的过程中积累了大量的图书借阅数据，形成特定的数据仓库。在这庞大的数据仓库蕴藏着许多可挖掘的重要价值。本文针对于高校图书馆，以广东外语外贸大学为例，分别分析以思科信息学院为代表的理工科和以英语教育学院为代表的文科学院 ，结合数据挖掘的知识，深入讨论其不同性质的专业学生群体的借阅行为，以及借阅兴趣取向。分析的内容主要从以下的几个方面展开了研究与探讨：

（1）基于时序模式分析学生借阅行为规律。

时序模式分析，是一种动态数据处理的统计方法，其中具有代表性为时间序列分析。用时序模式分析对学生借阅行为的发展趋势进行和预测，可以了解纸质图书馆借阅的发展情况、变化规律，未来行为等提供决策性帮助。

（2）基于实体对齐的学生借阅图书类别识别与统计。

学生借阅的图书类别或主题往往具有对应或主从关系。对图书类别单从标签分类必然导致较高的误别率。本文对学生图书类别识别基于实体对齐的思想，结合 WordNet 和维基百科、百度百科，利用半监督的方法，识别出学生借阅图书中的主流图书类别，以及类别之间的关系。

（3）基于图书标签的学生兴趣社区发现。

抽取学生借阅图书的标签，作为用户兴趣模型的特征，构建用户兴趣模型相似度网络。利用一种基于模块度增益识别社团的贪婪算法（BGLL）进行学生兴趣社区划分，并为对不同兴趣社区的特征，以及分布情况进行分析。

关键词：高校图书馆，数据挖掘，时序模式，实体对齐，社区发现。

目 录

第一章 绪论

1.1 研究背景与意义

高校图书馆每天积累了大量的图书借阅量，会产生许多的图书流通数据。目前针对这些数据，除了简单的业务统计，并没有进行深层次的专业的数据挖掘，因而数据的使用价值远远没有得到利用。通过大量的外借数据可以发现，读者和借阅图书之间存在一定的关联，不同专业背景的读者有不同的借阅习惯，产生不同的借阅行为。利用数据挖掘的知识，挖掘出这些数据的关联，能够提高资源的质量以及高校图书馆的服务水平，为高校图书馆的信息管理提供更好的帮助。

在图书馆管理中应用数据挖掘技术，有诸多的益处，一方面可以提高图书馆的管理质量；另一方面，可以针对图书馆读者的借阅信息进行分析，将隐藏在读者借阅行为中的信息、数据等发掘出来,为读者的图书借阅行为服务。信息技术的广泛应用促进图书馆服务工作的转变和改革。加上图书馆的图书信息、数据等非常多，管理工作的难度非常大，海天都有大量的图书被借阅或者是归还。在这个过程中图书馆的借阅服务管理工作发挥着重要的作用。使用数字挖掘技术可以对图书馆每天形成的大量的数据信息进行分析，自动的对图书借阅数据信息中隐藏的信息进行检索，针对读者的图书借阅需求进行预测发现读者的图书爱好为读者提供主动性的服务提高图书馆的服务质量。

随着社会的进步和发展，越来越多的信息技术应用在图书馆建设中将图书馆的服务质量提升，同时促进图书馆的信息化建设和数字化建设。数字挖掘技术在图书馆管理工作中的应用,帮助图书馆借阅管理人员对读者的借阅行为进行分析，以便于为其提供更好的借阅服务。数据挖掘技术的图书馆借阅管理中的应用可以优化图书馆的馆藏资源,将读者的借阅倾向反应出来准确的为读者提供借阅服务，可以说数字化技术在图书馆管理工作中肩非常好的应用前景。

1.2 高校图书馆的国内外研究现状

高校图书馆与一般图书馆有显著的区别，一般图书馆是面向大众化、受益群体可以为 3 岁至 90 岁，而高校图书馆针对于具有较高文化素养的学生群体。研究高校图书馆的价值，向来受到各界学者的青睐。在 CNKI 中以“高校图书馆&价值”作为主题词检索有 1900 多篇文献，相关度较高的文献也接近 200 篇。《高校图书馆的价值》一文是中国学者对美国 2010 年出版的一份关于高校图书馆价值的调研报告的编译和总结。该报告详尽地介绍了美国大学和研究图书馆协会（ACRL）通过调研得出高校图书馆价值定位结论的全过程。可见，研究和掌握高校图书馆的价值具有非常重要的意义。

研究高校图书馆学生的借阅行为情况，是研究高校图书馆的价值定向的一个重要方面。近几年有几篇文献通过专业的统计软件或者数据挖掘软件对图书馆中有数据记录的行为进行分析，主要集中于对借还书行为的分析。少部分文献开始对读者进行细分化研究。目的都是为了针对不同的读者，提供个性化的服务，增加图书馆的效用。

王伟（2012 年）运用 WEB 挖掘和书目挖掘技术对图书馆用户行为进行分析，选择关联规则、序列模式、分类模式、聚类模式以及时间序列模式对图书馆中的书目信息数据以及用户借阅数据进行分析。文章中没有实证数据的研究，只是引入挖掘技术，给出分析思路，对之后的学术研究具有启发和借鉴作用，但是由于没有实证研究，所以文章总体偏浅，对挖掘技术的可行性分析不到位，挖掘技术在应用中的问题也估计不足。

在 EBSCO 中的 LISTA 数据库中以“Library”&“user study”“behavior”等主题词进行检索，得到的文献数量不是很多。新南威尔士大学图书馆研究者在《CHANGES IN ACADEMIC LIBRARY SPACE: A CASE STUDY AT THE UNIVERSITY OF NEW SOUTH WALES》中分析了在图书馆实体重建前后用户的体验变化。主要采取的研究方式是问卷调查，对于来图书馆的人员通过问卷获知他们来图书馆的目的、对图书馆各项设备的使用情况以及满意度等。

《A study of academic library users' decision-making process: a Lens model approach》一文中作者通过问卷调查的模式对用户借阅行为的影响因子进行分析，发现用户喜欢运用各种他们能够获取的信息来帮助他们做借阅决策，因

此图书馆应该更加主动推送相关性高的信息，形成“提示”型的环境，在用户进行信息检索和信息获取的各个阶段给予用户指导性提示。

国外的用户行为分析文献中考虑的角度比较多，不但有实体图书馆发生的用户行为，以及数字图书馆中的信息行为，还有用户情感心理方面的行为。分析的模式多是借助各领域不同的模型对图书馆用户行为分析,实现跨学科领域的交叉研究。

1.3 数据挖掘的国内外研究现状

1997 发表的《数据挖掘技术》中第一次对数据挖掘做出明确定义：数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的新技术,它能挖掘出数据间潜在的模式，找出最有价值的信息，指导商业行为或辅助科学研究。随着时间的推移,技术不断发展更新,数据挖掘渗透的领域越之广泛,由原来的聚类,分类，关联规则等经典算法，演进成各个不同门派的领域，如文本挖掘、社会网络分析、机器学习等。在 Web of Science 中检索到的数据挖掘相关文献的年度分布情况看来，从 1997 年到 2011 的发文量大致可分为四个阶段：第一阶段：1997 年-2000 年，这是数据挖掘的起步阶段，发文量逐年上升，但是上升幅度较小；第二阶段：2001 年-2006 年，这是数据挖掘的发展阶段，该阶段的文献数量比上阶段有明显的增长；第三阶段：2007 年-2010 年，这是数据挖掘的过渡阶段，该阶段的文献数量先下降，再上升，最终回到 2006 年最高点时的文献量；第四阶段：2011 年之后，该阶段文献数量又呈现高速增长趋势。15 年里文献数量增长了 14 倍之多，在未来几年，预计数据挖掘的研究论文数量还会增长。

社会网络研究是数据挖掘中的一个子领域，而社团发现是的其中一个重要课题，是指通过某种规则将网络划分成多个子网络，使得在同一个子网络中节点间的联系较大，在不同子网络的节点间联系较小。识别出社会网络中的社团结构并对其进行分析是了解现实网络组织的一种重要的方法。社团识别在计算机科学、生物学和社会科学等领域都有广泛的应用。如通过分析社交网络中的社团结构可以得出人们的兴趣；分析通话网络可以得出用户的交往圈；分析生物网络可划分出癌变细胞社团等。

社团检测中最经典的算法是 Girven 和 Newman 在 2002 年提出的 GN 算法，

该算法根据网络中社区内部高内聚、社区之间低内聚的特点，逐步去除社区之间的边，取得相对内聚的社区结构。算法用边介数的概念来探测边的位置，某边的边介数定义为网络上所有顶点之间的最短路径通过该边的次数。

除此之外，Newman 在 2004 年提出模块度概念之后，Blondel 等在 2008 年提出了一种基于模块度增益识别社团的贪婪算法，该算法通过节点转移社区获得最大的模块度增益，逐步的获得最佳模块度，从而识别出社区结构。经试验表明该算法相比于其他算法，能在时间开销更小的情况下获得较好的社团识别结果。

1.4 课题研究的创新点

本文的主题是对高校图书馆学生借阅行为的分析研究，以数据挖掘作为研究工具，以广东外语外贸大学思科信息学院与英语教育学院的学生信息数据和学生借阅作为研究对象。课题的创新点在于以下方面：

（1）本文通过时间序列和图书类别、不同读者群体对比（文理对比，年级对比）三个维度的综合分析，研究不同读者群体在不同时间特性的借阅行为情况，分析读者借阅与归还的高峰期与低谷期，与相互时间差的背后原因。

（2）本文融合了命名实体识别与对齐的知识，加强了图书类别的分类效果。实验表明，该方法还能自发现不同类别之间的并列，或主从关系。

（3）本文运用社会发现算法，以学生兴趣相似度作为学生之间的实体关系，构建学生兴趣网络，自划分出不同社区群体。相比聚类算法，该算法的稳定性更高，划分结果更为明显。

第二章 数据的获取

2.1 网页爬虫技术

网页爬虫，又称为网页蜘蛛，是搜索引擎抓取网页系统不可或缺的组成部分。现网页爬虫成为了各领域获取数据来源的一个重要工具。网页爬虫技术的基本工作流程如下：

1.首先选取一部分挑选的种子URL；

2.将这些URL放入待抓取URL队列；

3.从待抓取URL队列中取出待抓取的URL，解析DNS，并且得到主机的ip，并将URL对应的网页下载下来，并利用正则表达式获取需要的数据信息；

4.整合数据信息到文件流或数据库中。

对于静态网页可以采取上述做法，而对于动态网页，需要用户登录认证的才能获取的网页。我们需要利用cookies技术。Cookie是http消息头中的一种属性，包括：

Cookie名字（Name）、Cookie的值（Value）

Cookie的过期时间（Expires/Max-Age),Cookie作用路径（Path）

Cookie所在域名（Domain），使用Cookie进行安全连接（Secure）

前两个参数是Cookie应用的必要条件。在动态网页爬虫技术，我们常采用会话Cookie（session cookie）：这个类型的cookie只在会话期间内有效，保存在浏览器的缓存之中，用户访问网站时，会话cookie被创建，当关闭浏览器的时候，它会被浏览器删除。

2.2 利用爬虫获取信息源

本文采用python语言实现爬虫框架，并利用cookie动态获取网页源代码，以爬取所需的网页元素。本文爬取的内容主要有：

（1）广东外语外贸大学电子图书馆，思科信息学院与英语教育学院14级、13级、12级的学生借阅记录；

（2）学生借阅书籍对应的索书号，主题；

（3）豆瓣图书网站的图书标签。

具体如图 1、图 2、图 3 所示：

No.	著者	题名	年	应还日期	应还时间	归还日期	归还时间	罚款	分馆
1	林翊	ERP综合实验教程	2013	20151204	22:00	20151123	11:08		南校区中文书外借分馆
2	赵雪松	ERP基础	2014	20151204	22:00	20151123	11:08		南校区中文书外借分馆
3	里杰斯	Java程序设计教程	2008	20151123	22:00	20151123	11:09		南校区中文书外借分馆
4	杰克逊	Android应用开发入门	2013	20150917	22:00	20151023	10:27		南校区中文书外借分馆
5	里杰斯	Java程序设计教程	2008	20150701	22:00	20150608	09:59		南校区中文书外借分馆
6	里杰斯	Java程序设计教程	2008	20150507	22:00	20150601	17:01		南校区中文书外借分馆
7	刘西杰	HTML、CSS、JavaScript网页制作从入门到精通	2013	20150504	22:00	20150421	18:47		南校区中文书外借分馆
8	庞永庆	21天学通Java	2009	20150409	22:00	20150421	18:47		南校区中文书外借分馆
9	科特勒	营销管理	2009	20150325	22:00	20150331	18:59		南校区中文书外借分馆
10	林树泽	Java完全自学手册	2009	20150325	22:00	20150311	10:00		南校区中文书外借分馆
11	金士发	炒股四要	2007	20150325	22:00	20150311	10:00		南校区中文书外借分馆
12	陈筱芳	人力资源管理	2008	20150325	22:00	20150311	10:00		南校区中文书外借分馆
13	吴德庆	管理经济学	2010	20141231	22:00	20141230	15:37		南校区中文书外借分馆
14	林树泽	Java完全自学手册	2009	20141230	22:00	20150106	20:41		南校区中文书外借分馆

图 1 广外电子图书馆学生借阅记录

选择显示格式: 标准格式 卡片格式 引文格式 字段名格式 MARC格式

格式	BK
ISBN	978-7-5141-2738-6 价格:CNY43.00 (含光盘)
	978-7-900276-32-2
作品语种	chi
题名	ERP综合实验教程
出版发行	北京 经济科学出版社 2013
载体形态	319页 : 图; 26cm+光盘1片
丛编项	经济与管理实验教学研究文库 jing ji yu guan li shi yan jiao xue yan jiu wen ku
一般性附注	本书是受中央财政支持地方高校发展专项项目"福建师范大学产业与区域经济综合竞技研究创新团队"资助的研究成果
丛编	经济与管理实验教学研究文库,
个人著者	林翊
全部馆藏	所有单册
分馆馆藏	北校区中文书外借分馆
分馆馆藏	南校区中文书外借分馆
可用馆藏	仅限被过滤的单册

选择显示格式: 标准格式 卡片格式 引文格式 字段名格式 MARC格式

图 2 学生借阅的书籍具体情况

作者简介 · · · · · ·

安迪·威尔 (Andy Weir)，从15岁起就被美国国家实验室聘为软件工程师。执着的太空宅男，沉迷于相对论物理、轨道力学和载人飞船。《火星救援》是他的处女作。

2009年，安迪·威尔陆续将他的小说《火星救援》贴在自己的个人网站上，供人免费阅读。在众多读者的强烈要求下，他在亚马逊平台上发布了作品，收费0.99美金，哪知花钱买他小说的读者比免费阅读的读者更多。2013年3月，兰登书屋以六位数买下小说的版权。仅仅四天后，安迪·威尔又接到了来自20世纪福克斯电影公司的橄榄枝。2015年，由大导演雷德利·斯科特执导、马特·达蒙主演的电影《火星救援》将于10月2日上映，更是激发了这本小说的购买热潮，直接将它推向了《纽约时报》畅销书榜的榜首位置。

豆瓣成员常用的标签(共214个) · · · · · ·

科幻　科幻小说　小说　美国　火星救援　安迪·威尔　外星求生　外国文学

图 3 豆瓣图书图书标签

图 1 为广外电子图书馆学生借阅记录，我们需要获取的学生借阅的书籍，书籍对应的作者，应还日期，实还日期；图 2 为学生借阅到的书籍的具体情况，获取的信息有书名，对应的 isbn，以及图书的主题；图 3 为豆瓣图书的图书标签，是标记图书类别信息的重要指标，如“科幻”，“科幻小说”等。

第三章 基于时序模式分析

3.1 时序模式的定义

时间序列是指按照时间顺序取得的一系列观测值，是随时间变化而又相互关联的数字序列,其观测值间具有依赖性。时间序列统计是从时间的发展变化角度，研究事物在不同时间上的一段时间内，其数量变化和时间的关系，探索其随时间推移的演变趋势和变化规律。通过时间序列的分解与组合，可进行季节因素及循环周期的测定和分析。在本文中，时序模式常被应用于读者的借阅时间的规律分析。如通过借阅记录的月、星期、日、小时的时间序列，可得到借阅周期和峰谷；通过对 n 年的平均一年中的月，平均星期中的日，平均每天中的小时的时间序列分析，可得到相应时间序列的借还规律。

3.2 数据集采样选取

本文研究的角度为学生借阅书籍的应还日期与实还日期，从数据爬虫可以获取每一位同学借阅书籍的具体应还时间与实还时间，数据精确度达到时分。但考虑到数据集本身的特性和时间片选取的合理性，本文将时间片以月为单位，统计每月中学生的归还数量和实还数量。

3.3 实验结果与分析

本实验以思科信息学院与英语教育学院 2015 年学生图书借阅情况作为研究对象，由于考虑学院中专业不同的影响因素，在思科信息学院的学生名单中只含有计算机专业学生（按行政班排列），除去信管专业和电子商务专业的学生名单。现思科信息学院 2012 级学生 318 人，2013 级学生 317 人，2014 级学生 351 人，而英语教育学院对应的人数分别为 101 人，107 人，104 人。图 1、图 2 分别为两个学院在 2015 年不同月份的应还数量和实还数量的对比情况：

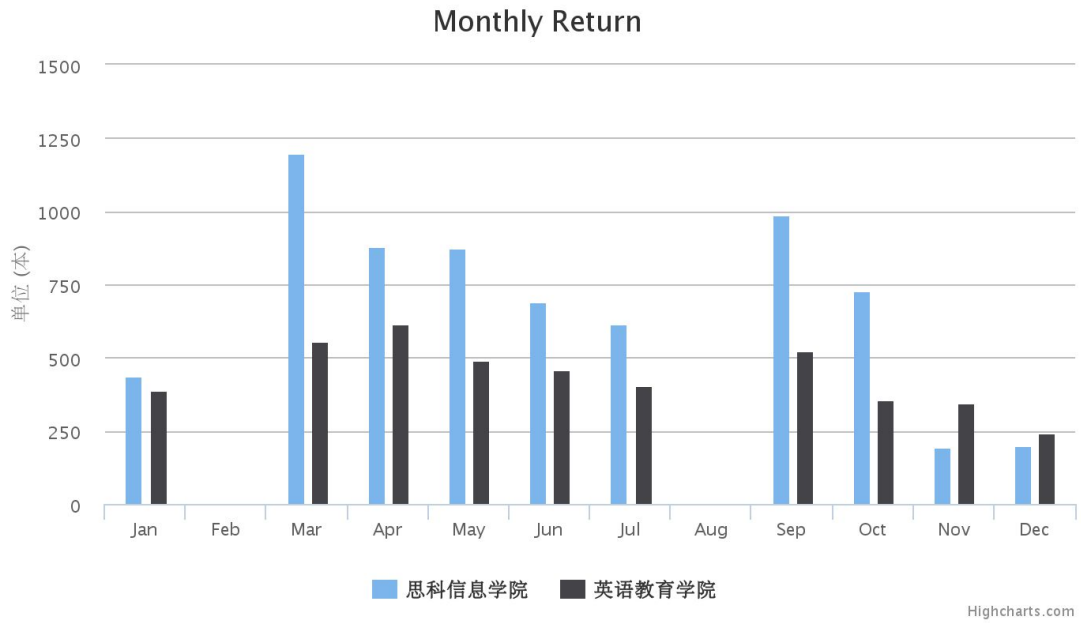


图 1 不同学院学生在 2015 年不同月份的书籍应还数量对比情况

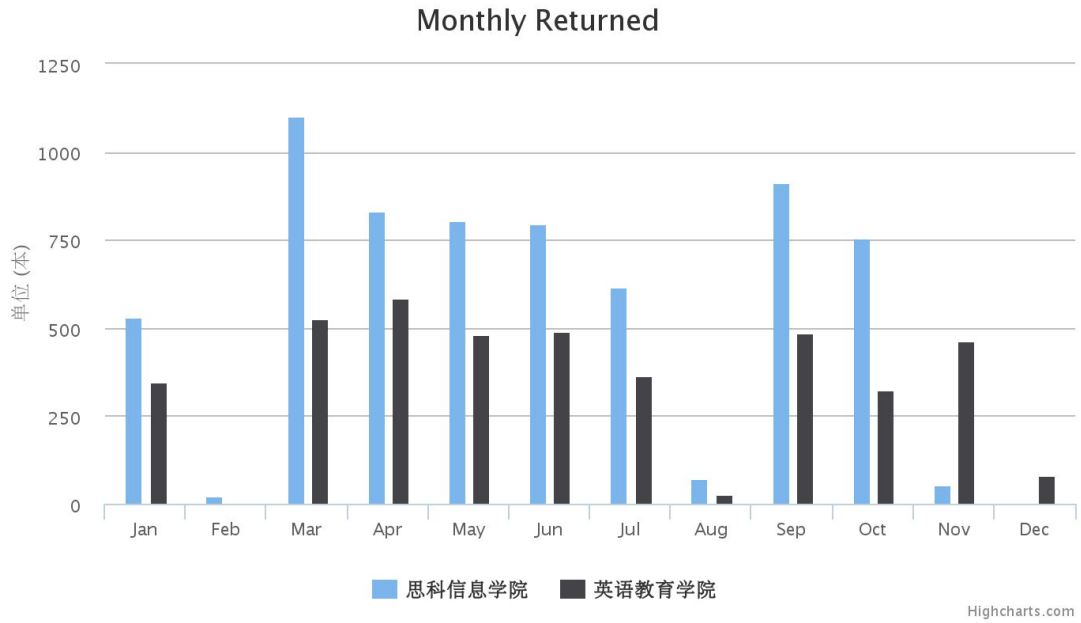
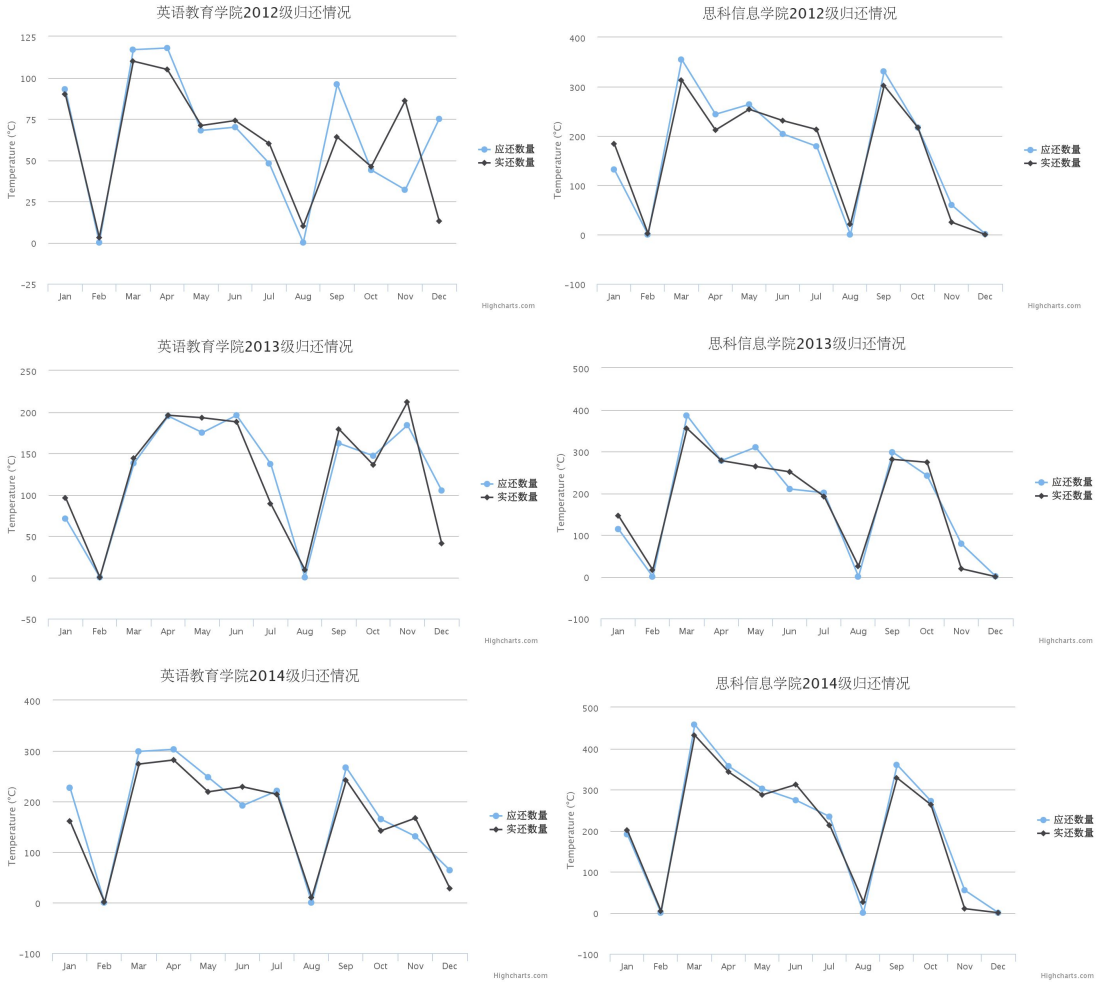


图 2 不同学院学生在 2015 年不同月份的书籍实还数量对比情况

从图表可以分析，两个学院的学生在 2 月份与 8 月份的图书借阅记录最少，在 3 月份与 9 月份的归还数量几乎为全年最高。但不同点在于，思科信息学院的学子借阅书籍的情况在第一学期是随学期的进展而呈递减式衰减，但英语教育学院的学子在第一学期的借阅书籍的峰值却集中在 4 月份；另外，从第二学期情况

看来，思科信息学院的学子普遍偏向于考完试而将图书还清，而英语教育学院的学子则大体偏于在考前清理借阅记录。从总体上而言，思科信息学院人均借阅图书为 6 本，英语教育学院人均借阅图书为 4 本。

除此，我们还针对不同年级的借阅情况做了以下分析：



由这些图表分析，我们可以清晰看到，除了英语教育学院 2012 级学生以外，其余学院和年级的应还情况与归还情况基本一致，而不同专业的借阅情况由明显的区别，尤其在于下半年，英语教育学院呈双峰值，而思科信息学院则层阶梯式递减，这是由于 11 月份中有全国英语笔译等级考试，英语教育学生在这一时期借阅英语类书籍比较多。从细的来看，英语教育学院 2012 级学生，在 9 月和 11 月的借阅情况产生了巨大的反常，而思科信息学院 2012 级学生的借阅情况十分稳定和正常；两个学院的 2014 级学生整体上，应还数量在于实还数量之上，很可能出现较多的拖欠想象。

第四章 基于实体对齐的图书类别分析

4.1 图书类别分类

图书类别分类是属于文本分类的中一个应用。将图书视为文本的一种形式，提取图书的内容或标签为图书的文本特征。构建文本特征映射，从而利用分类器对图书进行分类。分类器代表有朴素贝叶斯分类器，SVM 分类器等。但针对于本文提供的数据集情况，文本分类的思路存在以下的弊端：

传统的文本分类，类别之间只存在相容或排斥的关系，忽略了类别与类别的主从关系，如 python 属于计算机编程语言类等。图书类别存在较强的类别主从关系。传统的分类算法显然不适用。

另外，从数据集的情况来看，学生借阅图书类别广泛且借阅数量不一，导致分类的维度增加，出现较多的离群点情况。

总而言之，从文本分类的角度思考对该实验的数据集进行图书类别分类不是一个理想的方法。为解决上述的不足，本文引用基于实体对齐的思想，为解决图书类别多维度的情况，采用类似一趟聚类的方法对图书类别进行分类。从实验来看，图书分类的较为理想，能够识别大部分类别之间的主从关系。

4.2 基于实体对齐的图书类别关系识别

实体对齐，是针对一个实体有多种表达方式的情况下，寻求其对应的关系。在图书类别中，常常存在类别之间的主从关系，如小说与文学存在主从关系，如史铁生与文学存在主从关系。本文将图书在图书馆对应的主题以及在豆瓣图书对应的社会化标签作为该图书的主要特征，利用一趟聚类的思想以及实体对齐的方法，分类的具体步骤如下：

（1）以 isbn 作为图书的唯一编号，图书馆的作者、名字、主题、社会化标签都作为图书标签；

（2）对数据进行清洗，选取频数出现较高的标签作为图书的主标签。

（3）任意选取一个图书，选取其所有主标签，形成 n 个主题；

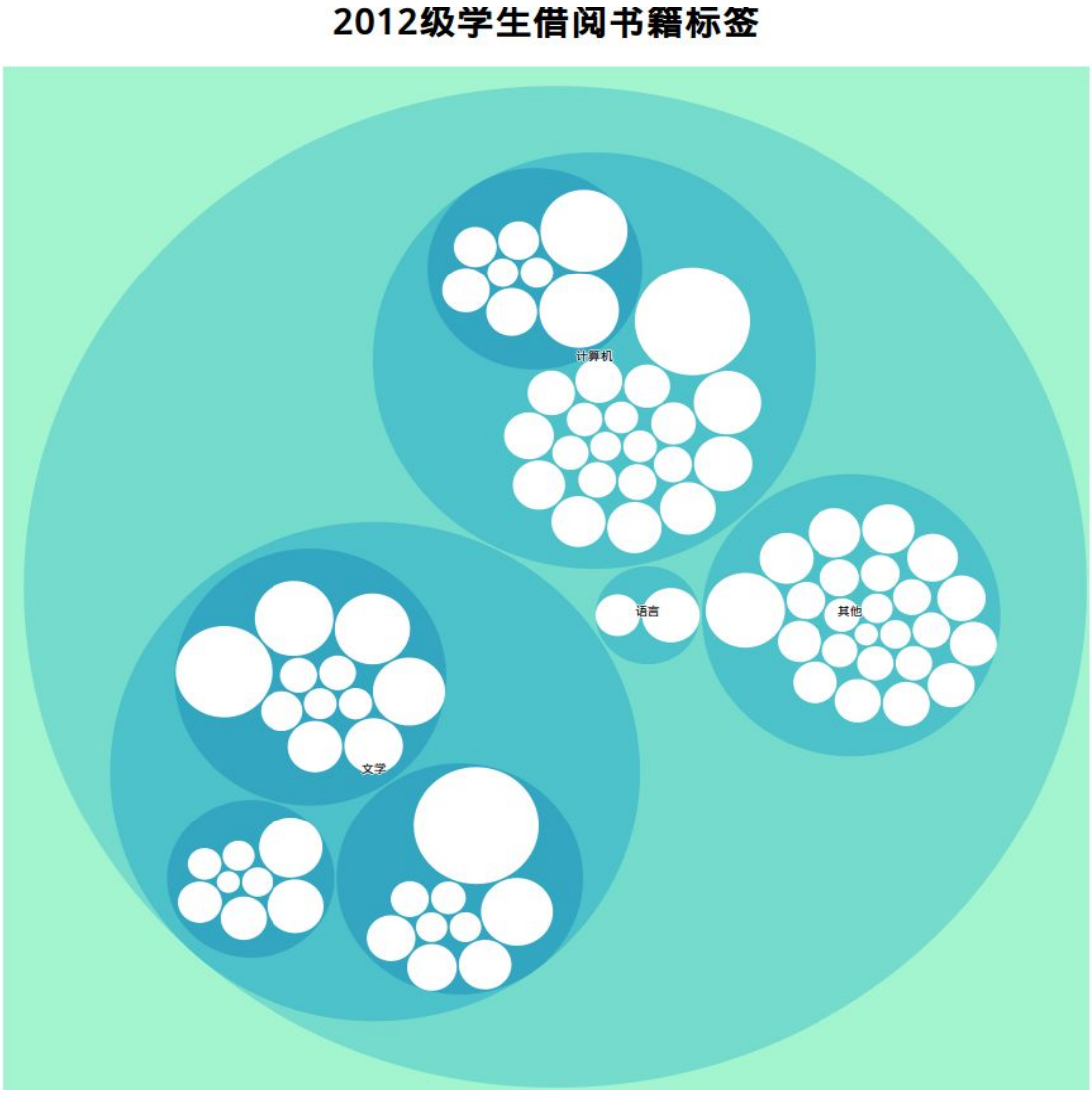
（4）利用爬虫技术在维基百科和百度百科寻找主题之间的主从关系，若有，

则小主题归并大主题当中，若无，则形成新的主题；

（5）重复第（3）步骤，直至所有图书遍历完毕。

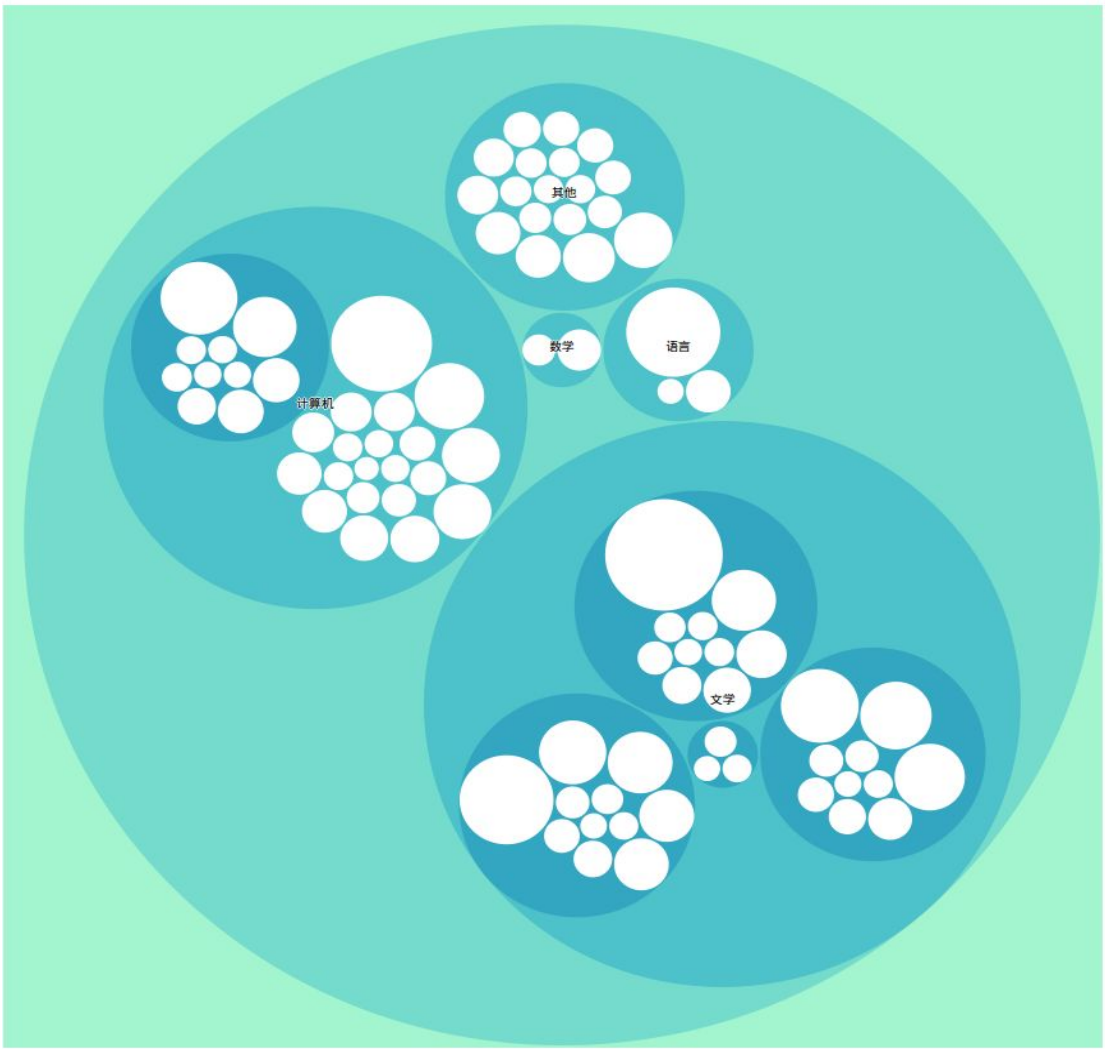
（6）对最后的结果进行人工的监督和过滤。

我们通过用思科信息学院分别对 12 级学生、13 级学生和 14 级学生的借阅图书进行分类。由于分类的类别比较多 ，我们选取所有图书中主要的类别。分类的结果如下：



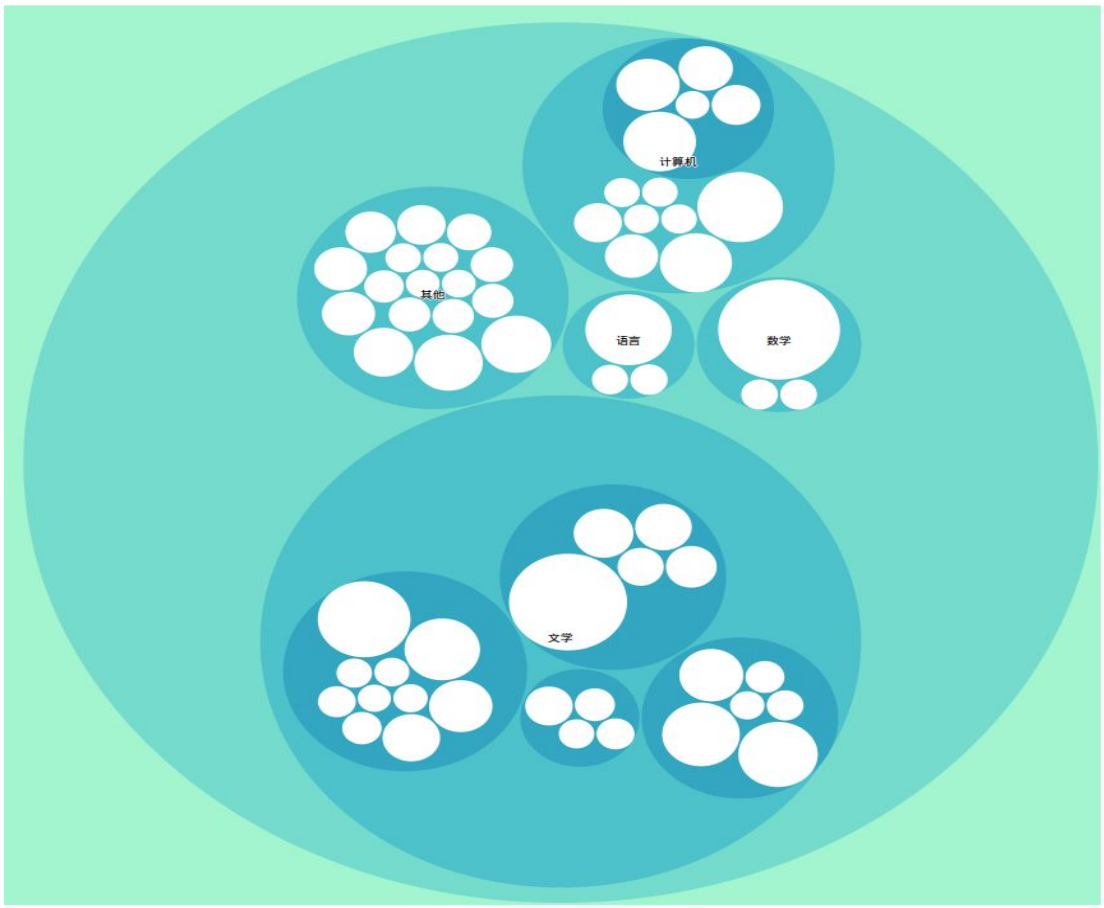
2012 级的思科信息学院学生主要的类别有计算机、文学、语言。由于 12 级学生有 3 年的借阅历史，所以借阅的书籍最为广泛、种类最多，将类别较多的都归为其他类，其中包括随笔、经济、管理、艺术、励志、科普等；在计算机类中，计算机语言和编程居多，而 java 语言与 c 语言高居榜首；文学类以小说为主，小说中由以长篇小说为主，而历史主题在文学中较为显著。

2013级学生借阅书籍标签



2013 级学生的情况除了计算机、文学、语言作为主流外，增加了数学，而数学以高数、线性代数为主；在文学类中，13 级学生出现了较多哲学类的图书，分别有哲学史、人生哲学、西方哲学等；计算机类依然以编程为主，但计算机语言类别相对丰富，且分布均匀，如出现了 python，c++，android，php 语言等；在其他类别中，心理类占的比例最重。语言以英语为主，日语、韩语次之。

2014级学生借阅书籍标签



2014 级学生中，文学类的主题占的比例最大，出现了中国文学、经典文学、日本文学、美国文学、法国文学等多类型的文学体裁。而计算机以数据结构和编程为主，计算机语言种类相对较小；数学以高数为主，新增了离散数学。其他类别中，以随笔和心理为主。

4.3 实验总结

从三个年级来看，思科信息学院的学生都以计算机、文学、语言类为主。其中文学占的比例要比计算机类的要大，而总体来看，学生借阅计算机相关类的书籍在总借阅书籍中超不过 50%。对于广东外语外贸大学这一文科类院校，思科信息学院作为该校唯一的工科类专业，以及加上双学位的影响，出现这一现象实属正常。不过本章节的重点在于利用实体对齐的方法寻找的类别之间的主从关系，据实验的结果来看，该方法对图书馆图书类别分类起到良好的作用。

第五章 基于图书标签的学生图书兴趣社区发现

5.1 社区发现

社区发现是社会网络研究的其中一个重要课题，是指通过某种规则将网络划分成多个子网络，使得在同一个子网络中节点间的联系较大，在不同子网络的节点间联系较小。识别出社会网络中的社团结构并对其进行分析是了解现实网络组织的一种重要的方法。

社区发现中最经典的算法是 Girven 和 Newman 在 2002 年提出的 GN 算法，该算法根据网络中社区内部高内聚、社区之间低内聚的特点，逐步去除社区之间的边，取得相对内聚的社区结构。算法用边介数的概念来探测边的位置，某边的边介数定义为网络上所有顶点之间的最短路径通过该边的次数。

除此之外，Newman 在 2004 年提出模块度概念之后，Blondel 等在 2008 年提出了一种基于模块度增益识别社团的贪婪算法，该算法通过节点转移社区获得最大的模块度增益，逐步的获得最佳模块度，从而识别出社区结构。经试验表明该算法相比于其他算法，能在时间开销更小的情况下获得较好的社团识别结果。

Donetti 等人在 2004 年提出了一种基于谱聚类的社团识别算法，该算法假设相同社区中的拉普拉斯矩阵的特征向量具有相似性，然后将向量装换成在矩阵空间中上的一个点，最后对矩阵空间中的点进行聚类，划分社团。

Hughes 的随机游走算法也能识别出社团，Zhou 等人采用随机游走的方法计算节点间的距离，并假设距离更近的节点集更像是属于同一个社团。Palla 等人在 Nature 上发表了一个重叠社团识别的算法。该算法假设网络中有 k 个节点全耦合的子网络构成，并称为 k-派系。如果两个 k - 派系有 k-1 个公共节点，那么就称这两个 k - 派系是相邻的。如果一个 k - 派系可以通过若干个相邻的 k - 派系到达另一个 k - 派系，就称这两个 k - 派系为彼此连通的。网络中由所有彼此连通的 k - 派系构成的集合就称为一个 k - 派系社区。

本次实验，考虑到基于模块度增益识别社团的贪婪算法具有较高的稳定性，采用算法对学生借阅图书的兴趣社区发现。

5.2 基于模块度增益识别社团的贪婪算法

该算法主要分为两个步骤：第一步，从节点合并开始，构建第一步社团划分结果。每个节点根据模块度增益决定是否加入到邻居节点的社区中和到底加入到哪个邻居节点的社区中。每个节点按序执行该过程。(2) 重新构建网络。把第一步每个社团单做一个节点，边是原来的社区之间链接边权的和。迭代 (1)，(2) 直到收敛。

具体如下：

第一阶段，首先进行社团初始化，网络中的每个节点都分配一个社团编号，这样每个节点都被看作是一个社团。然后，对于任意节点 i，j ，根据式(2)计算当节点 i 加入到它的每个邻居节点 j 所在的社团时，对应社团模块度的增量ΔQ：

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

式中 ∑in 是社团内部所有边的权重和；∑tot 是所有与社团内部节点相关联的边的权重和；k_i 是所有与点 i 相关联的边的权重和；i,k_{in} 是节点 i 与社团 C 相连接的所有边的权重和。 当ΔQ 为正值时，选出对应最大值的那个邻居节点，把点 i 加入到该邻居节点所在的社团中；若所有ΔQ 都为负值，则节点 i 留在初始社团中。这种社团的合并过程重复进行，直到整个网络不再出现合并现象时，划分出了第一层的社团。

第二阶段，首先构造一个新网络，该新网络的节点是第一阶段探测出的各个社团，节点之间连边的权重是两个社团之间所有连边的权重和。然后，用第一阶段中的算法再次对该新网络进行社团划分，得到第二层的社团结构。以此类推，直到不能再划分出更高一层的社团结构为止。基于模块度增益识别社团的贪婪算法本身就能够生成一种层次性的社团结构。

社团模块度 Q 是当今比较流行的评价社团识别效果的指标。模块度越高，则表示社团识别效果越好，社团内部更紧凑。其定义为：

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j)$$

对于非加权无向网络， m 为网络中连边的数量；A_{ij}为网络中节点 i 到节点 j 的连边； k_i和k_j分别为与节点 i 和节点 j 有关联的连边数量；当 c_i =c_j时(c_i，c_j）=1 反之则(c_i，c_j）=0.对于加权无向网络 m 为网络连接的权重之和； A_{ij}为网络中节点 i 到节点 j 的连边权重；k_i和k_j分别为与节点 i 和节点 j 有关联的连边权重。

对于有向网络中，模块度定义为：

$$Q = \frac{1}{W}\sum_{i,j}[W_{ij} - \frac{S_i^{out}S_j^{in}}{W}]\delta(c_i, c_j)$$

对于非加权有向网络， w 为网络中连边的数量； w_{ij}为网络中节点 i 到节点 j 的连边； s_i^{out}和s_j^{in}分别为节点 i 作为起始节点和节点 j 作为终止节点的连边数量；当 c_i=c_j时(c_i，c_j）=1，反之则(c_i，c_j）=0 。对于加权有向网络 w 为网络连接的权重之和； w_{ij}为网络中节点 i 到节点 j 的连边权重；s_i^{out}和s_j^{in}分别为节点 i 作为起始节点和节点 j 作为终止节点连边权重之和。

5.3 实验分析与对比

本实验的对象分别是思科信息学院与英语英语教育学院 12 级，13 级，14 级的学生，收集学生图书借阅数据，以图书记录作为学生的兴趣趋向，构建学生兴趣度增广网络，利用基于模块度增益识别社团的贪婪算法，得到不同网络划分后的社区情况。具体的步骤如下：

（1）根据学生借阅书籍的情况，构建学生的特征文档，文档内容有学生借阅图书的名字，主题，作者，社会化标签，并过滤一些不重要的信息（通过分词与词性识别，过滤掉无意义的程度副词）；

（2）利用 python 的结巴分词技术，将文档细分为词条标签；

（3）根据词条标签，运用余弦相似度的方法，计算学生兴趣的相似度，构建学生兴趣相似度增广网络；

（4）根据增广网络，调用 python 的 igraph 包，运用 BGLL 算法对增广网络进行社区划分。

划分结果如下：

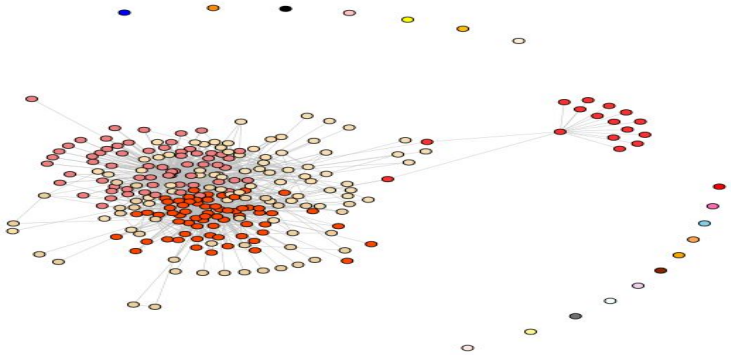


图 3 思科信息学院 2012 级学生图书兴趣社区发现

如图 3 为思科信息学院 2012 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.9，划分的社区个数共有 24 个，社区模块度为 0.301.

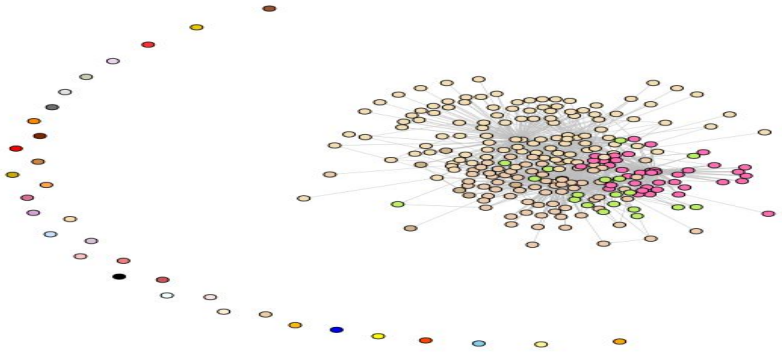


图 4 思科信息学院 2013 级学生图书兴趣社区发现

图 4 为思科信息学院 2013 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.88，划分的社区个数共有 38 个，社区模块度为 0.314.

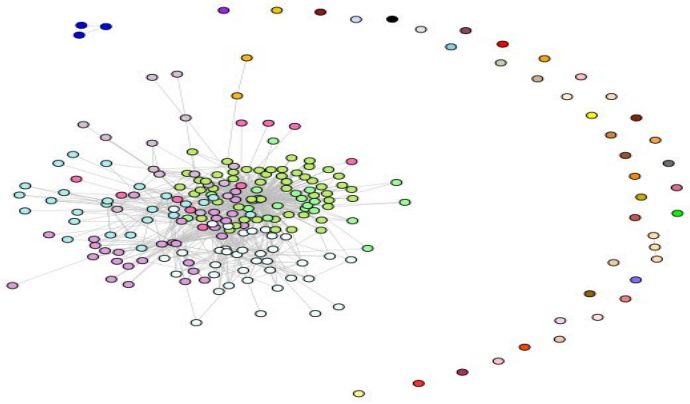


图 5 思科信息学院 2014 级学生图书兴趣社区发现

图 5 为思科信息学院 2014 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.85，划分的社区个数共有 48 个，社区模块度为 0.363.

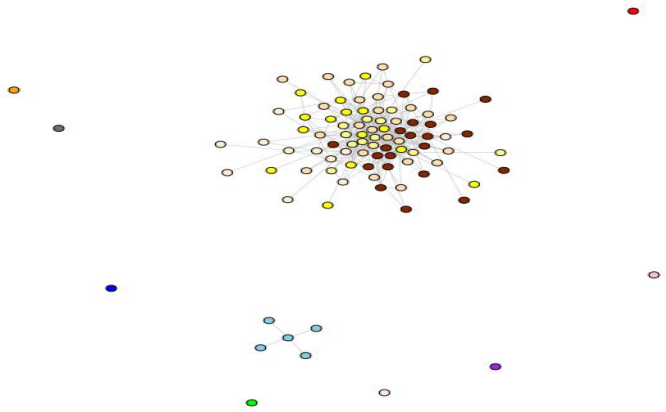


图 6 英语教育学院 2012 级学生图书兴趣社区发现

图 6 为英语教育学院 2012 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.77，划分的社区个数共有 14 个，社区模块度为 0.353.

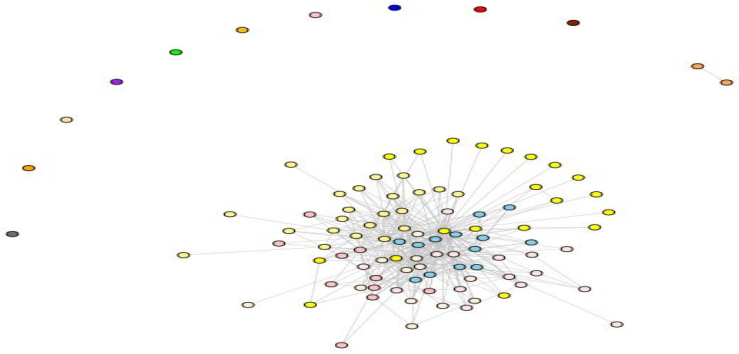


图 7 英语教育学院 2013 级学生图书兴趣社区发现

图 7 为英语教育学院 2013 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.77，划分的社区个数共有 15 个，社区模块度为 0.294.

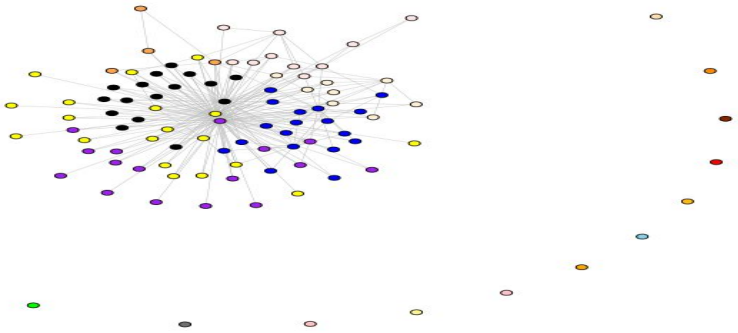


图 8 英语教育学院 2014 级学生图书兴趣社区发现

图 8 为英语教育学院 2014 级学生图书兴趣社区发现情况，该网络中学生兴趣度阈值为 0.80，划分的社区个数共有 17 个，社区模块度为 0.339.

从划分的结果来看，思科信息学院的学生孤立点情况比英语教育学院的要多很多。追踪结果发现，思科信息学院的学生的兴趣点相对与英语教育学院的广泛很多，前者借阅的类型有文学类、管理类、经济类、计算机类、科普类等，而后者，较为集中于英语语言类与文学类，而且外文文献居多。思科信息学院的学生基本以文学类、英语类、计算机类为主要社区，而且学生借阅类型较多较，需要较高的相似度才能区分为不同的社区。而英语教育学院的学生则群体不大，学生的借阅类型不多，相似度不需那么高就能区分开，但从社区模块度来看，英语教育学生划分结果没有思科信息学院学生的要理想。

5.4 实验改进方案与实施

在上一次实验中发现，两个学院社区划分的结果都出现较多的离群点。分析其中的原因，发现学生的兴趣并不是一成不变，随着时间的推移和人生阅历的增长而发生改变，所以将学生的兴趣视为静止的时间片来分析，具有一定的误差。针对这一不足，我们讨论，提出如下的方案：

将学生的借阅书籍记录按着时间先后顺序追加到学生文档中，利用莱文斯坦距离计算学生之间的兴趣相似度，改进原来的兴趣相似度增广网络。

我们对出现较多离群点的思科信息学院学生兴趣社区，以上述的方案，重新实验。实验结果如下：

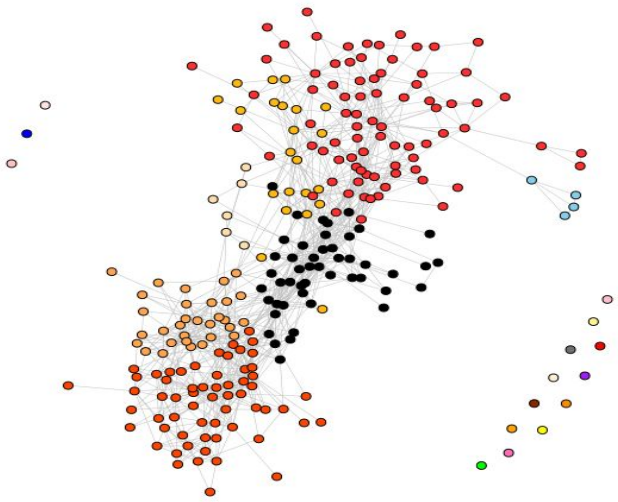


图 9 改进后的思科信息学院 2012 级学生图书兴趣社区发现

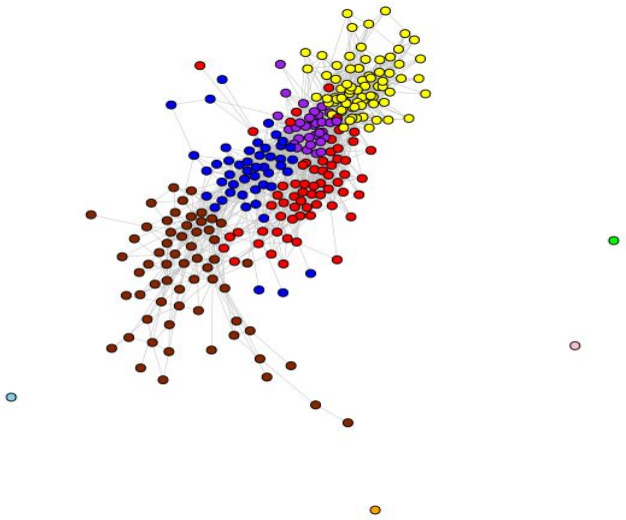


图 10 改进后的思科信息学院 2013 级学生图书兴趣社区发现

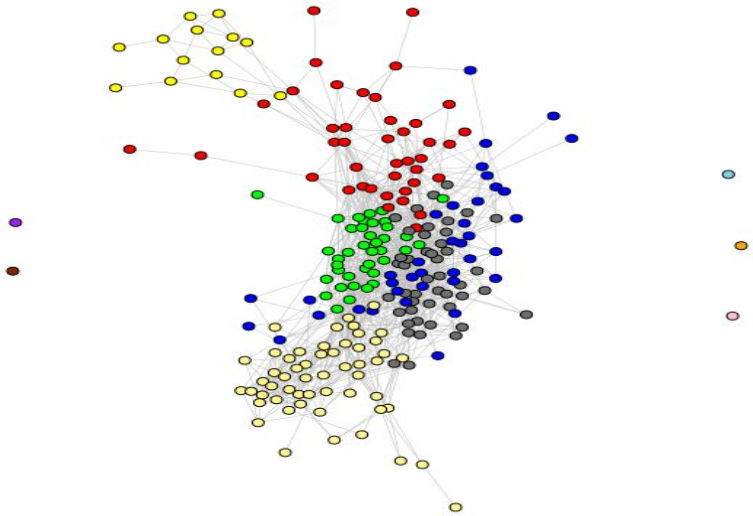


图 11 改进后的思科信息学院 2014 级学生图书兴趣社区发现

由图 09，图 10，图 11 可以观察，孤立点的个数明显减少。该实验中，学生的兴趣相似度阈值由原来的 0.90，0.88，0.85,分别降为 0.30，0.31，0.30，而社区模块度分别提高至 0.547，0.4952，0.467.社区分布也相对理想，从图中看来，区分度也提高很多。

第六章 调查分析

6.1 问卷设计

针对本报告的分析结果，为了挖掘其背后的原因，我们设计了一份调查问卷，调查的对象主要是广东外语外贸大学思科信息学院与英语教育教育学院大二至大四的学生，具体的问卷如下：

1. 您的年级？［单选题］［必答题］

○ 2012 级（大四）

○ 2013 级（大三）

○ 2014 级（大二）

○ 2015 级（大一）

2. 您所在的学院是？［填空题］［必答题］

_____

3. 您平均月借阅书籍数量？［单选题］［必答题］

○ 5 本以下

○ 6-10 本

○ 11-15 本

○ 16 本以上

4. 您平时借阅的图书类型是？［多选题］［必答题］

□ 英语类

□ 其他语言类

□ 计算机类

□ 小说类

□ 经济管理类

□ 哲学政治类

□ 科普类

□ 心灵鸡汤类

□ 其他

5. 您借阅书籍的目的是？［多选题］[必答题]

☐ 兴趣出发

☐ 考试需求

☐ 帮别人借阅

☐ 作为工具书

☐ 其他

6. 您归还书籍的拖欠情况如何？［单选题］[必答题]

◯ 从不拖欠

◯ 一个月 1-2 次

◯ 一个月 3-4 次

◯ 一个月 5 次以上

◯ 很少借阅书籍故无拖欠

7. 您借阅的书籍与您专业是否相关？［单选题］[必答题]

◯ 非常相关

◯ 较相关

◯ 一般相关

◯ 很少相关

◯ 从不相关

8. 您借阅书籍的峰值集中于哪个时期？［单选题］[必答题]

◯ 开学初

◯ 学期中

◯ 临考前

◯ 考试结束后

9. 您觉得您对图书的兴趣类别经常变更吗？［单选题］[必答题]

◯ 从不变更

◯ 基本稳定

◯ 与之前稍有变更

◯ 经常变更

10. 您觉得变更的原因是？［多选题］［必答题］

☐ 课程安排原因

☐ 阅历增长导致看法不同

☐ 偶发事件改变兴趣

☐ 他人疏导

☐ 其他

6.2 问卷统计与分析

参与本次问卷调查的人数情况为：思科信息学院 17 人，英语教育学院 36 人。在调查中，人数主要集中在大二上，如 12 所示。

选项	比例
2012 级（大四）	16.95%
2013 级（大三）	27.12%
2014 级（大二）	55.93%

图 12 参与调查学生的年级情况

在平均月借阅书籍上来看，在之前的分析中，思科信息学院的人均月借阅书籍是 6 本，而英语教育学院是 4 本。但抽样调查中，两个学院的情况均在 5 本以下，两个学院只有 7 人的是在 6-10 本之内。这方面间接反映了两个学院的借阅情况在月季中呈明显的波峰状，导致整年的平均借阅书籍与调查情况偏高。具体如图 13 所示：

选项	比例
5 本以下	86.44%
6-10 本	11.86%
11-15 本	1.69%
16 本以上	0%

图 13 人均月借阅书籍情况

在拖欠的情况中分析，思科信息学院的出现拖欠情况在 17 份中有 4 份出现拖欠，而英语教育学院 36 份中有 7 份出现拖欠情况，但据此，难以分析出思科信息学院的与英语教育学院的拖欠情况的对比。

而在借阅图书类别中，思科信息学院的学生出现类别整体比英语教育学院的学生较多，思科信息学院学生借阅类别分别有小说类、哲学政治类、科普类、计算机类、英语类、经济管理类，而英语教育学院的学生主要集中在英语类与小说类、心灵鸡汤类之间。不过英语类、小说类在两个学院中都居列榜首。图 14 为整体类别的情况：

选项	比例
英语类	54.24%
其他语言类	5.08%
计算机类	20.34%
小说类	66.1%
经济管理类	11.86%
哲学政治类	18.64%
科普类	11.86%
心灵鸡汤类	11.86%
其他	23.73%

图 14 两个学院共同借阅图书类别的整体情况

分析到学生借阅图书的目的，两个学院学生大部分上都是以兴趣为出发点，主要体现在借阅小说类之上。但两者存在一定的区别，对专业书籍的借阅上，英语教育学院的学生偏向于考试需求为驱动，而思科信息学院的学生则以将其作为工具书为目的。图 15 为学生借阅图书目的统计情况。另外，前者借阅书籍体现较强的专业相关，后者则相对较弱。据此，我们通过咨询的方式，了解到思科信息学院学生借阅书籍专业相关性较弱的主要原因：

一、思科信息学院学生对计算机专业的热爱程度较低；

二、计算机专业知识日新月异，图书馆的书籍相对陈旧，学生偏向于购买自行购买，或利用网络资源。

选项	比例
兴趣出发	83.05%
考试需求	49.15%
帮别人借阅	8.47%
作为工具书	52.54%
其他	8.47%

图 15 学生借阅图书的目的情况统计图

对于兴趣类别的变更，学生表示基本稳定，而调查变更的缘由，主要有课程安排的原因、阅历增长导致看法不同、偶发事件改变兴趣，第二占的比例相对较大；而其余两者的比例不相伯仲。这反映了学生的兴趣会随着时间的推移发生一定的变化，而这一变化是一个稳定的过程。图 16、17 分别为学生兴趣变化，以及变化原因的统计情况。

选项	比例
从不变更	3.39%
基本稳定	76.27%
与之前稍有变更	13.56%
经常变更	6.78%

图 16 学生兴趣变更的统计情况

选项	比例
课程安排原因	40.68%
阅历增长导致看法不同	55.93%
偶发事件改变兴趣	52.54%
他人疏导	13.56%
其他	6.78%

图 17 学生兴趣变更原因的统计情况

第七章 总结与展望

6.1 成员分工情况

本报告主要成员由张礼明、林家骏、赵彦青、孙博轩四人。报告主题的提出人为张礼明，并负责统筹整个报告的进展情况，提供技术指导。另外，林家骏负责可视化模块展示，包括柱状图、折线图、图书类别分布图等；赵彦青负责收集学生信息，编辑文档、设计调查问卷并统计等；孙博轩负责收集文献材料、网页爬取，并为数据进行简单统计。成员分工明确，报告按期完成。

成员评分 ：张礼明（90）、赵彦青（85）、林家骏（80）、孙博轩（80）

6.2 对未来的展望

数据挖掘技术现已广泛应用于各行各业，尤为金融、电信等领域，并取得了成功。数据挖掘的应用不仅促进了社会经济快速发展、改善了人民的生活质量，而且产生了深远的影响。本次项目将数据挖掘技术应用于图书馆领域，通过对图书借还记录的挖掘，对数据进行深层剖析，通过现象看本质，为图书馆管理提供指导依据，为读者开展更加亲民化的服务，改变图书馆传统的服务模式，为学生们提供更舒适、更方便的服务，提升广外图书馆的服务质量。

本项目在研究了数据挖掘技术的发展、应用、过程和算法后，将数据挖掘技术用用到图书馆的管理中，以广东外语外贸大学思科信息学院计算机系和英语教育学院的大二到大四学生的借阅数据为例，对数据进行挖掘，不过，在之后的研究中仍需完善以下几方面：

1、增大覆盖范围

本次数据只覆盖到思科信息学院计算机系和英语教育学院的学生，所以数据基础还是比较薄弱的，在之后的研究中，应该广泛覆盖到广东外语外贸大学更多学院大一到大四的学生借阅数据，对数据经过挖掘之后，分析结果更具说服力，改善服务更全面。

2、数据的更新

数据挖掘是一个延续性的过程，所以数据及时更新是非常重要的，每次挖掘结果只是针对每次选择的数据，随着数据的增加应及时更新，并通过实践应用不

断检验和完善。

3、加强对挖掘结果的研究

随着科学技术的快速发展，挖掘算法也是日新月异并且越来越有效，所以之后的研究可将精力集中在挖掘结果的研究和验证上，结果越多可能出现的情况更多种多样，甚至可能超出我们个人理解范围，所以我们要对结果进行筛选，使结果更简单易懂。

数据挖掘技术在图书馆的应用还处于初步阶段，但其中数据的价值远远没有完全实现，相信在未来，随着个性化、亲民化服务理念的不断深入，数据挖掘对图书馆的发展能起到更好、更有效的推动作用。

参考文献

[1]付沙，《基于序列模式挖掘的图书馆用户借阅行为分析》，情报理论与实践，2014，06；

[2]何 琳，万 健，何 娟 ，郭诗云，《基于社会标签的中文图书自动分类研究》，数字图书馆，2014，09；

[3]王菲，《数据挖掘在图书馆用户行为分析上的应用研究》，上海交通大学硕士论文，2013，06；

[4]陈文文，《图书馆使用者行为模式的数据挖掘研究》，2007，06；

[5]李艳，《高校图书馆读者借阅行为分析》，科技情报开发与经济学报，2013，05；

[6]王娜，李霞，徐红英，《社会网络分析之社区发现研究》，深圳大学学报，2014，01；

[7]蔡晓妍，戴冠中，杨黎斌《基于谱聚类的复杂网络社团发现算法》，计算机学报，2009，09；