分类号：_____ 密级：公开

UDC _____ 学号：2016022292

華南師範大學

South China Normal University

硕 士 学 位 论 文

(学术学位)

基于关键词网络的学者推荐研究

学位申请人: 付海林

专业学位名称: 软件工程

专业学位领域: 数据挖掘

所在院系: 计算机学院

导师姓名及职称: 李建国

2019 年 1 月 15 日

基于关键词网络的学者推荐研究

专业名称： 软件工程
申请者： 付海林
导师： 李建国

摘 要

互联网技术迅速发展，各行各业的数据都在指数膨胀，科研领域也一样，数据量急速的增长给科研人员带来了丰富的信息，同时也带来了许多难题。作为科研学者，需要实时保持追踪学术的新动态，然而，通过人工从海量数据中筛选出与自己研究领域相符的学者，无疑是一个非常浩大的工程。因此，本文使用自然语言处理方面的技术快速提取文本的关键词，再结合图论相关算法，快速筛选出与用户学者研究兴趣相符的科研学者，进而推荐给学者用户，以帮助科研人员节省时间，减小工作量。

本文提出了一个基于关键词的学者推荐模型，该模型包含挖掘论文学者研究领域（关键词提取）、用户学者兴趣（关键词网络）和打分推荐三个子模块。主要流程是通过融合多种特征从论文摘要中提取出关键词集合，将关键词集作为论文学者研究领域；接着，将关键词集和用户学者搜索的关键词结合，构建以共现为关系的词图 KOG，通过多种算法挖掘 KOG 图得到该图的核心节点，本文将得到的核心节点作为对用户学者偏好或者兴趣的一个扩展；最后，构建论文作者（Author）- 关键词（Keyword）的二部图 AKG，来对论文学者用户打分排序，其中关键词出现在论文学者的摘要中，作者节点和关键词节点之间就存在边相连。本文充分利用关键词去挖掘用户学者 - 论文学者之间的隐式联系。

本文之所以关键词作为切入点，是因为短小精悍的关键词不仅能够代表一篇论文的主题，而且其本身蕴含着丰富的信息，其被广泛的应用在文本分类、聚类、

搜索、推荐等领域。本文只考虑使用关键词（不使用任何其它的特征，例如论文引用数、作者影响力、用户学者的基本信息或者历史记录等）实现学者推荐，所以本文对于解决推荐系统中的冷启动问题有很重要的作用。无论是从用户学者的角度还是从论文学者的角度考虑，本模型都围绕关键词展开，关键词是影响该学者推荐模型准确率的唯一因素，因此，关键词抽取问题即是本文研究的核心部分，本文提出了融合多特征的关键词抽取算法，并对结果进行评估。

本文在 KDD、WWW 真实的论文数据集进行关键词提取的对比实验。为了验证本文推荐框架的有效性，本文爬取了 Microsoft Academic 官网的生物医学论文数据进行实验验证，实验证明，该推荐模型到不错的准确率，能进行有效的推荐，对于解决大数据时代信息过载问题，本文的研究具有重要的实际意义。

关键词：学者推荐；关键词抽取；词网络挖掘；核心点选取；冷启动

A scholar recommendation framework based on keyword

Major： Software Engineering

Name： Hailin Fu

Supervisor： Jianguo Li

ABSTRACT

KEY WORDS: scholar recommendation; extract keyword; cold start

目 录

第 1 章　绪论

1.1　研究背景与意义

随着时代的进步, 互联网、物联网、云计算、三网融合等 IT 与通信技术的迅猛发展，大数据成为信息技术领域的又一热门话题，在数据挖掘、人工智能、社会计算、生物学和化学等领域的应用日渐深入，信息社会已经进入了网络大数据时代。在当前网络信息时代的背景下，可以说大数据时代已经到来，而且这个新概念被赋予了极为丰富的内涵[?]。近几年，大数据越来越显示出巨大的影响作用，已逐渐应用于政治、医疗健康、学术科研、教育、图书馆服务等领域。据统计，国外知名社交网站 Facebook 管理超过 400 亿张图片，所需的存储空间超过 100PB，每天发布的消息高于 60 亿条，需 10TB 空间存储，Twitter 每天产生 1.9 亿条微博；搜索引擎每天的日志超过 35TB，Google 搜索引擎一天需要处理的数据超过 25PB。2010 年全世界的信息总量是 1ZB，最近 3 年人类产生的信息量已经超过历史上总信息之和，预计到 2020 年，总量将达到 35ZB[??]。大数据既给人们的生活带来了便利，同时也带来了巨大的挑战，该技术正在改变着人们的工作与生活方式[?]。

与此同时，自 2007 年起，国内外出现了很多专业的科研社交网站，如国外的 ResearchGate、Academic 国内有科研之友、学者网、百度学术等，为学者提供了线上的交流平台，科研交流有了新的有效途径[?]。在学术科研领域中，学术文献也以飞快的速度在增长，每天数以万计学术成果被发表，1997 年据 R.M.May 的统计，公开出版物的年增长速率为 3.7%，特别是在一些比较热门的研究领域，现在这个数字更为惊人[?]。韩增义也在其论文中表示代表人类知识前沿的科技文献正在以每年 6%-8% 的速率增长，不同国家或地区之间合作论文的比例从 2003 年的 13.2% 增加至 2013 年的 19.2%，2013 年 Scopus 数据库收录的同行评议期刊所发表的论文数量达 219.97 万篇[?]。

面对如此庞大的数据，作为科研学者，要追踪所在学科的最新发展动向，就要追踪该领域的相关文献，分析这些文献中发表的新观点，用到的新技术，得到最新的研究热点以及新的成果，最后追踪到其研究突出的相关学者。然而，新兴

技术往往是几个学科相互交叉的结果，而从一个学科去发现别的学科中出现的新技术、新主题或者相关学者就更加困难，如何快速有效地获取与自身研究相关的学者将是一个很有挑战的问题。目前，虽然已经有很多学者对社交网络开展了广泛研究，但多数是在探索该类网站对线上科研交流与学术创新的促进作用，特别是在国内，科研社交网络的研究大多是探讨国内外发展差距、调查用户需求等，国内科研社交网络的应用现状还远落后于国外，网站中的资源丰富性、内容可获得性、检索结果可用性等都差强人意。

本文的研究内容是结合自然语言处理和推荐系统相关理论知识，实现快速向用户推荐可能感兴趣的学者。充分利用文献摘要的重要性，提取文本的重要特征，构建推荐对象的兴趣模型，再结合用户搜索的关键词信息，构建相关模型来识别用户的搜索意图，综合考虑推荐对象和用户两者之间的特征信息挖掘两者之间的隐式联系，从而发掘出用户感兴趣的学者，以此来帮助科研学者快速的找到有相同研究兴趣的学者。本文对科研社交网站中的学者推荐进行研究，有利于增强学术合作、提升科研人员学术交流、提高科研学者的工作效率和推动国内的科学研究都具有深远意义。

1.2 国内外研究现状

大数据时代，既给人们的生活带来了很多便利，同时也面临着很多的问题。你可以足不出户便轻松了解到所有的大小事情，同时大量的信息也阻碍我们获取高质量的信息。面对信息膨胀的问题，各个领域的研究者也都已经提出了解决办法，例如分类目录、信息检索和推荐，毫无疑问推荐系统在解决信息过载问题上是非常成功，自从推荐系统被提出之日起，便吸引了广大研究者的注意，而且被广泛的应用在各个领域中，例如电影推荐[? ?]、音乐推荐[? ? ?]、图书推荐[?]、广告推荐[?]、电商推荐[?] 以及学术领域推荐。本文的研究领域则是学术领域相关的推荐，科研社交网站作为一种新型的专业社交网络平台，主要关注的是为科研人员提供在线的以科学研究为导向的活动及构建学者间学术网络[?]。从 2007 年开始，国内外就出现一些科研社交相关的网站，如国内外的 ResearchGate、Academic 和科研之友和学者网等等，为科研交流有了新的有效途径。

经研究表明，在对文献数据进行挖掘时，论文正文篇幅较长，且包含的冗余信息较多，最能代表论文的内容的是论文的题目、摘要和关键词等部分[?]，关键词短小精悍，但其包含的信息却非常的丰富，因此本文就采用文本挖掘和自然语言处理相关技术抽取出论文摘要中的信息作为关键词，以抽取出的关键词作为论文作者的兴趣。在此基础上提出基于关键词网络的学者推荐模型，这里的关键词即是通过文本特征工程相关技术从文本中抽取出来的能代表作者研究兴趣的一些词或者短语；学者推荐则是属于社交网络范畴，因此本文将从特征提取之关键词抽取和社交网络两个维度去阐述目前的研究现状。

首先对关键词提取进行调研得知，关键词的提取是文本挖掘的一个子领域，而文本挖掘技术又是数据挖掘的一个分支，所以关键词抽取也就属于数据挖掘领域范畴。1995 年 Feldman 提出了文本挖掘概念和框架。国外研究比较早，在特征工程、文本分类等方面都取得丰富的研究成果。国内 1998 年开始才陆续开展文本挖掘的研究，并且由于中文自身的特点，难度系数也相对大，所以跟国外相比还存在着一定差距[?]。该技术在数据指数增长的时代扮演者重要的作用，很多文本挖掘工具都已经应用在商业方面[?]。通过将非结构化的文本转化为结构化数据的形式，让计算机能够计算，从而抽取出隐含的、有用的知识。文本特征工程的技术包括预处理、特征提取等，数据的预处理对结果准确率有很大的影响，是一个很重要的环节。该项技术被广泛应用在文本分类、聚类和情感分析等各项研究中。其中文献[?]就是采用将文本挖掘和深度学习结合对文本进行情感分析。50 年代末期，国外学者.P.Luh 就已经提出了词频统计的思想，用于自动分类，随后众多学者在该领域也取得卓越的成效，最近研究者主要围绕文本的挖掘模型、特征抽取和文本表示[?]。国内起步较晚，针对中文信息处理还未形成完整的技术理论和框架，不过进展也在逐步加快。

1.3 研究内容和方法

1.3.1 研究内容

本文通过调研分析科研社交网络的发展状况后，分析了目前科研社交网络在国内外理论和实际应用的情况，本文提出一种新的思路去解决学者推荐模型，详

细阐述了该学者推荐模型的工作步骤和方式，受文本特征提取之关键词抽取的启发，本文还提出了结果评价的指标。最后本文通过爬取“微软学术”官网的真实数据对提出的推荐模型的有效性进行验证。本文总体分为六个章节，具体每章的内容安排如下所示：

1.3.2 研究方法

本文主要采用的研究方法有如下几点：第一，文献调研法。通过互联网技术访问线上各个数据库中检索了大量研究领域的相关书籍、论文等学术成果，经过对国内外的相关研究文献与资料的全方位收集和分析，确立本文研究方向和主题，设计本文推荐模型框架及各模块之间的耦合。第二，迁移法，本文是建立在自然语言处理、文本挖掘、推荐系统等相关技术的研究基础之上，通过综合探索以上技术理论，将其迁移到本文的模型框架及设计的评价指标上。第三，实验仿真与分析法，本文通过调研和分析大量文献，提出本文的研究内容，为了验证本文提出的模型的有效性，本文在多个真实的数据上进行了实验，从而检验模型的可靠性。

1.4 本文主要贡献

本文通过大量前期调研，对比国内外科研社交网站在学者推荐技术上进行的研究，通过深入探索后提出了本文的研究问题，并针对提出的问题进行了大规模的对比实验，直到得出最后的结论。整个过程中本文的主要贡献体现出如下几点：

1. 本文提出了一个基于关键词网络的学者推荐模型，该模型能够根据新用户搜索的关键词进行及时推荐。该模型包含两个网络图，即关键词共现图 (Keywords Co-occurrence Graph)，为了描述方便，本文简称该图为 *KCG*，*KCG* 是根据用户学者检索的关键词和匹配到的文献的关键词共同构建的，其目的有二，第一是在用户学者没有明确的意图的情况下，通过挖掘 *KCG* 中的核心点作为用户学者的搜索意图，第二是通过该图确立用户学者的研究兴趣。另外一个是论文学者 (Author) 和关键词 (Keyword) 构建的二部图 (Graph), 简称为 *AKG*，该图的目的是采用某种算法对论文学者进行打分排序，以关键词为纽带，建立论文学者和用户学者之间的映射关系，从而向用

户学者推荐最有可能感兴趣的论文学者。本文设计的推荐模型通过耦合并挖掘 *KCG* 和 *AKG* 两个网络图，最后完成最终的推荐目标。

2. 本文设计多层过滤器从每篇摘要中抽取关键词集，抽取出的关键词即为论文学者的研究领域或者兴趣。多层过滤器包括采用自然语言处理相关的 Pos-Tag 标注词性、过滤停用词、正则匹配等启发式的算法过滤生成关键词候选集，还包括文本特征提取中的多个经典特征算法对候选集做进一步过滤，最后采用聚合排序算法对候选关键词集进行重排序，从而生成高质量的关键词集。

3. 在学者推荐领域，本文提出了全新的假设，即以关键词网络作为切入点，完成推荐任务。该假设是在缺乏其他指标，如行为记录、基本信息和论文引用数、影响因子等特征，仅仅只使用关键词这一个特征的条件下，设计出了本文的学者推荐模型，该模型能有效解决推荐系统的冷启动问题。为了验证有效性，本文爬取了微软学术官网的文献数据，在真实的数据上验证了模型的可行性。

第 2 章　相关理论

2.1　推荐系统

推荐系统的定义和概念很多，但 1997 年 Resnick 和 Varian 给出的定义被广泛接受，既“推荐系统是利用电子商务网站向客户提供商品信息和建议，帮助用户决定应该购买什么产品，模拟销售人员帮助客户完成购买过程”。

通用的推荐系统模型如图2-1所示，图中可以清晰看到，推荐系统中很重要的 3 个模块分别是：用户建模模块、推荐对象建模模块推荐算法模块，它通过匹配用户模型中兴趣需求信息和推荐对象模型中的特征信息，同时结合使用相关的推荐算法进行筛选，推荐用户可能感兴趣的对象[?]。

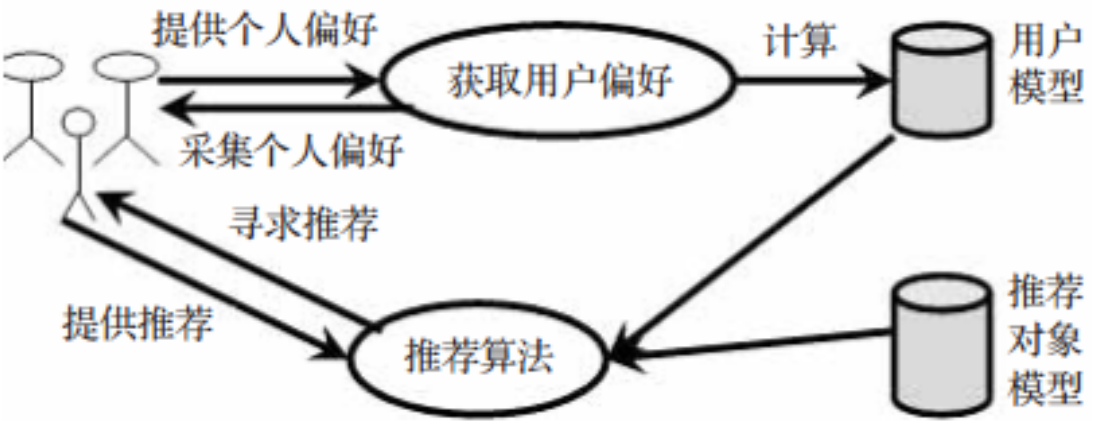

图 2-1　推荐系统的通用模型

2.1.1　个性化推荐概述

个性化推荐能成功，需要具备两个条件，第一是海量的信息，因为只有信息量很大的时候用户才需要系统的自动推荐。第二是用户没有明确的需求，如果有明确的需求，用户都会选择通过浏览器搜索，快速发现感兴趣的东西，而不需要推荐了。随着互联网行业的快速发展，信息量的增长速度飞快，各种新闻的推送(微博、微信公众号等)，占用了很多的时间，严重的影响了获取信息的质量问题，大量的垃圾信息导致人们获取有价值的信息的成本有所增加，并且，人们的生活

节奏也日益加快，需要在有限的时间内获取对自己有用的信息便难上加难，为了解决信息过载的难题，研究人员边通过用户历史行为数据开始对用户兴趣进行建模，从而实现个性化推荐的功能，让每个用户都有不一样的个性化页面。个性化推荐系统的价值便在于此。至今，国内外许多大型的公司投入了大量的精力到推荐系统的研究中，因此也给公司带来了很大收益，如最早研究推荐系统的亚马逊。

在学术社交研究领域个性化人物推荐包括社交好友推荐，论文合作作者推荐等。收集用户的历史行为记录，如评分、分享、收藏、自定义标签等，通过分析用户的行为，给其推荐可能感兴趣的对象，以及用户的个人基本信息，如性别、年龄，在推荐系统的研究中，真正难以把握的是用户个性化需求，

2.1.2 推荐算法

2.1.3 推荐系统存在的问题

冷启动问题推荐系统需要数据作为支撑，通常需要根据用户历史行为记录去预测未来可能产生的行为和兴趣。现实中，我们面对大量的新用户又或者对于全新上线的系统不仅没有有效的用户行为数据而且也缺乏用户的个人基本信息 (年龄、性别等)，我们没有任何数据对用户的偏好进行建模，这个问题被称为“冷启动”，是推荐系统中面临的一个难题问题。在缺乏用户行为数据的情况下，并不是就没办法给用户推荐，早期的推荐系统会基于商品内容数据做推荐，例如给商品打大量的标签或者通过推荐热门商品等方式来解决冷启动问题。Yao[?] 等学者提出通过用户搜索的关键字信息去预测用户的偏好，具体方式是通过比较搜索的关键词和已有的产品的信息做相似性比较，实验证明，他们提出的方法比推荐最受欢迎的方法效果更好。

噪音问题大数据时代面临一个很严重的问题，不仅数据量多，而且数据不干净要获得高质量的数据往往是不太现实的，需要前期花费人力物力进行数据预处理，高效的数据处理能力，能快速的挖掘数据的重要信息。随着需求的提出，研究者们也提出了很多数据处理的方式，例如减噪、归一化等。

数据长尾问题个性化推荐系统一开始主要解决的问题就是通过发掘长尾数据来提升商家产品的销售额，而长尾商品却仅仅只能代表小部分用户的需求，所

以只能充分发掘用户的行为，研究用户的兴趣，以此找到用户的个性化需求。所以，推荐系统可以更好地发掘数据的长尾。

2.2 关键词抽取

关键词抽取是从文本中提取出重要的短语或词。关键词背后所蕴含的丰富信息可以高度概括一篇文章的主题信息，该问题一直是自然语言处理中基础问题，也是研究的热点问题，关键词的研究成果被广泛的应用在文本分类、聚类、文本摘要、话题监测、索引和搜索等领域[????]。早于 20 世纪 80 年代已经有研究者用 TF-IDF 进行关键词的抽取。随着研究不断的推荐，近年来研究者们不断的提出新的特征或算法，如 florescu[?] 提出计算词出现的平均位置，caragea[?] 引入引文的特征等。对未来的性能提升方面，常耀成表示可以综合利用现有特征或提出新特征[?]。赵京胜等人通过从各个维度综述前人的工作后也提出了关键词抽取人物未来的发展方向[?]。Hasan[?] 通过全面调研总结了目前研究关键词抽取人物准确率普遍很低的原因，研究者也可以通过解决这些存在的问题作为研究思路。

2.2.1 关键词抽取流程

2.2.2 候选关键词特征

2.2.3 关键词抽取方法

关键词抽取主要被分为两个大方向，即有监督和无监督。由于有监督学习需要大量的人工标注，因此，在实际应用中，无监督的方式用得更加广泛。有监督算法[?] 提出的基于 SVM 分类器的有监督的关键词抽取，caragea[?] 提出基于引文网络的特征，在无监督算法中，florescu[?] 提出了基于位置特征改进 TextRank 的无监督自动关键词抽取算法。

2.2.3.1 无监督抽取

2.2.3.2 有监督抽取

在有监督的研究上常常把关键词抽取任务看成二分类问题，对候选关键词进行分类，即是关键词或不是关键词。首先建立分类器，分类器可以是朴素贝叶

斯[?]、决策树[?] 逻辑回归等。建好分类器后再用大量已经标注好的语料去训练分类器，再用训练好的分类器完成信息的自动抽取。在有监督问题中，通常被选用的特征有词性、词频、词位置 (首次出现位置、平均位置、最后出现的位置等) 和外部知识库等。不同的特征适用于不同的分类器。

，如 k-core、信息熵

2.3 学术社交网络

学术社交网络能为具有共同兴趣的科研工作者提供一个实时沟通、共享成果的平台。目前，学术社交网络的发展很迅速，其覆盖的面也非常广泛，功能也逐渐强大起来，为科研工作者提供很多科研社交服务。随着推荐技术的成熟和发展，在系统中加入推荐功能成为社交网络的热点，可以通过潜在好友关系进行人物推荐，也可以根据相似兴趣推荐书籍或论文等。实时为科研工作者推送一些推荐条目，不仅可以节约科研工作者的时间成本，好的用户体验，还可以吸引更多的新用户共享研究成果。有相关研究者对国外 12 个社交网络使用情况进行了调研，参加调研的人中超过 3000 科学家或工程师表示他们知道这些大型的设计网站，但是仅仅只有不到一半的人会定期的去访问 ResearchGate。详细信息如图2-2所示，该图摘自 2014 年 Richard Van Noorden 在自然上发表的文章[?]，从图中可以看出 Google Scholar 学术社交网站被定期访问的人数是最多的。针对学术社交网站的特殊性文中还对用户使用 ResearchGate、Academia.edu 和 Mendeley 三个社交平台的日常功能进行了调研，详细如下图2-3所示。国内的社交网站也非常多，例如微博、微信、知乎、百度学术、科研之友、学者网等，但是，本文暂时没有检索到类似的统计相数据。

社交网络的发展时间脉络如图，引自**??**

为了对比国内外学术社交网络发展的状况，接下来，本文着重介绍国内外四个知名的科研社交网络平台：

ResearchGate[1]被称为“科学研究的脸书 (facebook for research)”，自 2008 年上线到现在，该平台的研究者数量已经高达 15 亿多，包括 45 名诺贝尔获得者，学

[1]https://www.researchgate.net

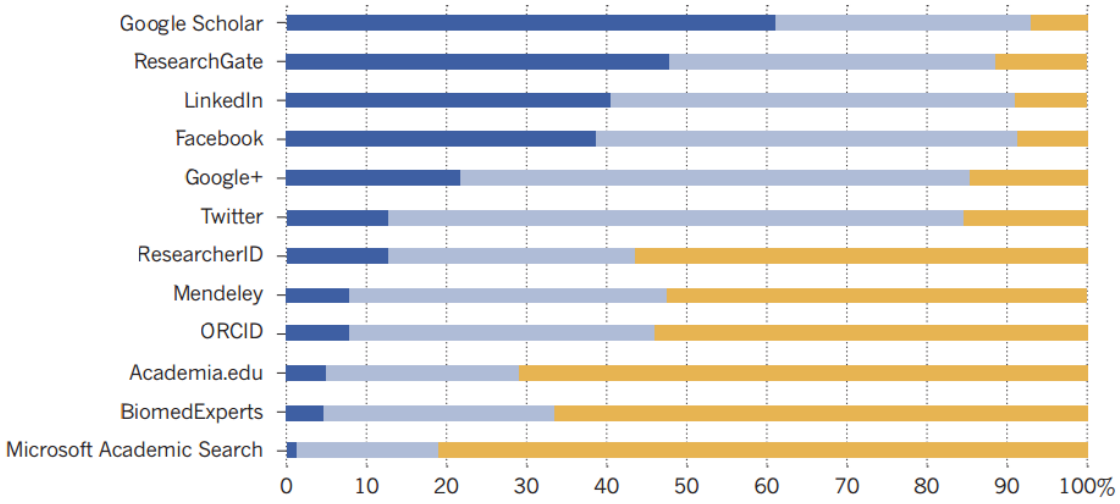

图 2-2　国外社交网络使用情况

注解：深蓝色表示知道有社交网站并且会定期访问的占比，浅蓝色表示知道有这些社交网站但是不会定期访问的占比，黄色表示不知道有这些科研网站的比例。

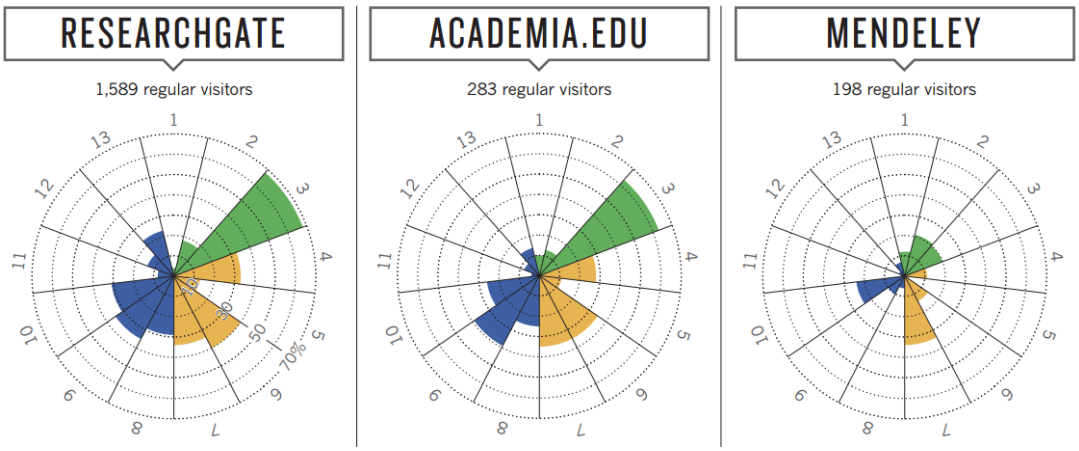

图 2-3　调研定期访问学术科研网站的学者的目的

注解：1 表示不常用，2 表示出于好奇，3 表示可能有学者联系，4 表示追踪指标，5 表示找工作，6 表示找合作伙伴，7 表示查看推荐的论文，8 表示联系同行，9 表示发布研究内容，10 表示分享链接，11 表示发起相关研究的讨论，12 表示对研究进行评论，13 表示跟踪讨论进展；其中有 1589 位常访问 ResearchGate 的用户参加了调研，283 位参加了 Academia.edu 调研，198 位参加了 Mendeley 的调研。

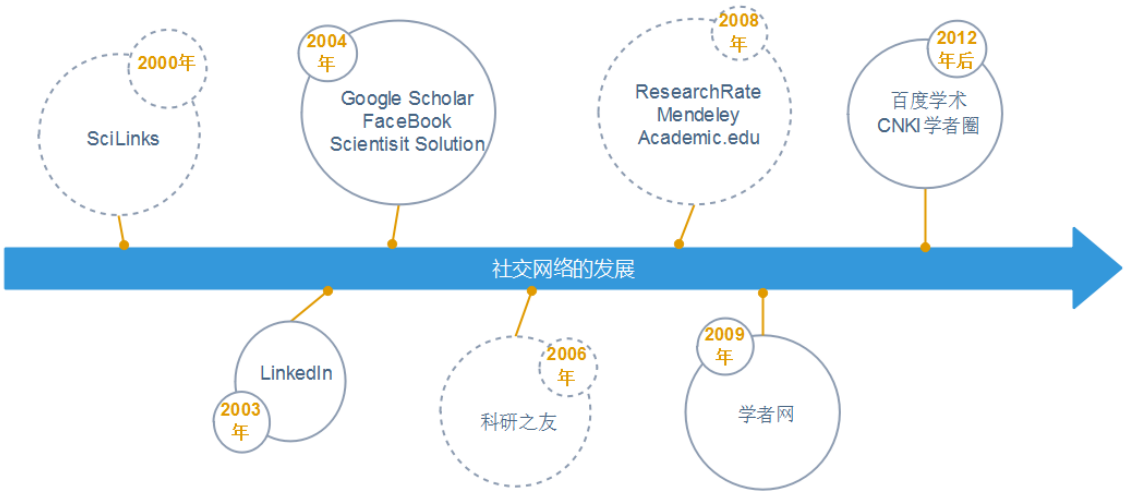

图 2-4　社交网络的发展

者可以在此网站上分享研究成果、学术论著、并且也可以加入一些科研论坛小组。通过注册时填写感兴趣的领域、专业知识等信息，其他跟你有相似研究兴趣的科研工作者可以很容易发现你。根据个人兴趣，ResearchGate 会推荐有相似研究兴趣的研究成果，此外，该网站还会根据用户兴趣推荐工作，

Academia.edu[2]于 2008 创建, 截止到 2018 年 12 月，该网站统计已经有 71 亿学者加入 Academia.edu 科研社交网站中，上传了 21 亿论文，并且每月的访客超过 40 亿，是一个非常庞大的数字。发表在 PLOS ONE 上的一篇文章表示，上传到 Academia.edu 的论文的 5 年内的引用率可以提高原来的 69%。

学者网[3]也

Microsoft Academic[4] **Google Scholar**[5] **百度学术**[6]

2.4　本章小结

[2]https://www.Academia.edu
[3]http://www.scholat.com/
[4]https://academic.microsoft.com
[5]https://academic.microsoft.com
[6]https://academic.microsoft.com

第 3 章　基于关键词网络的学者推荐模型

学术社交网络的目标是建立一个能发现用户兴趣的模型，一方面用作者已发表的论文的摘要去探索作者的研究领域，将具有共同兴趣的人联系起来，这样既可以提高已有用户的忠诚度，又有助于吸引新的用户，以此提升网络的流量。另一方面通过挖掘关键词共现网络，获取用户学者的潜在研究兴趣，匹配两者的兴趣点，将有相似潜研究兴趣的论文作者推荐给用户学者。准确的推荐功能不仅能提高用户的参与度，同样也可以提高社交网络的影响力。在社交网络中，通常会倾向于将有相似兴趣的人建立联系，因此对人物关系和兴趣定位建模是两个很重要的部分。本文构建关键词网络和论文作者 - 关键词网络，并将两人物关系网结合起来，可以更好的预测出与用户学者更相似的论文作者, 从而使本文的推荐性能更好。

3.1　学者推荐模型

3.1.1　目标问题定义

本文设计的学者推荐模型的目标是给无任何历史记录和个人资料的情况下，仅仅获取用户学者检索时输入的关键词，在文献数据的基本上实现学者推荐，该文献数据包括论文的标题、论文的摘要、论文的作者三个字段。通过用户检索关键词建立用户学者和论文作者之间的隐式联系，根据发现的联系给用户学者推荐其可能感兴趣的论文作者。举个简单的应用场景，对于一个新服务应用上线后，除了能搜集到用户的关键词外，基本无法获得一个游客用户的任何信息，针对游客怎么进行推荐？这个问题在推荐系统中被称为冷启动问题，本文以该问题为切入点，进行深入的研究，提出了下文的学者推荐模型，该模型在解决冷启动问题上非常优秀。并且该模型具有很强的移植性，能够在各式各样的场景中应用，例如将该模型应用在学术科研领域，不仅可以给学者用户带来很多好处，同时也可以给系统研发方本身也带来一定的益处。对于学者而言，该模型能够以少量的信息发挥无限的可能，快速的及时帮助用户学者挖掘到其感兴趣的领域学者。对于

系统而言，该模型成功应用在科研社交应用中能够提高用户学者的用户体验，从而吸引旧用户再次回访或者新的用户注册成为长期用户，以达到提高用户的留存率、网站点击率及其它效果。

关键词的高度抽象既是用户和作者的研究兴趣，因此，本推荐模型通过挖掘用户学者搜索的关键词和作者发表的论文摘要，找到该用户学者和论文学者之间的潜在联系。通过耦合关键词网络 (*AKG*) 和论文作者 - 关键词网络 (*AKG*)，发掘用户学者和论文学者的共同兴趣。

3.1.2 模型介绍

3.2 关键词网络

文档由许多词语特征组成，关键词对文档具有一定表征能力，因此，从与用户检索匹配到的文献摘要中抽取出关键词，这些关键词在一定意义上用户学者检索的内容，另一角度也反应了论文学者的研究内容，因此，本文的关键词网络是基于词项共现 (Term Co-ocurrence) 构建的，核心思想是用词项之间的共现程度反映语义之间的联系，经过多种方式挖掘关键词网络，获得将用户学者和论文作者之间联系较为紧密的词项，以此得到了两者之间联系的桥梁。

在没有推荐系统的时期，一个科研学者想投身于一个全新的研究领域时，一开始他只简单的了解了该领域的概况，对于该领域的具体分支以及对哪个分支的研究内容感兴趣，因此，他会通过关键词检索，检索并查阅大量的文献，从而确定自己的研究内容。例如张三想研究自然语言处理 (NLP) 领域，但自然语言处理包含很多任务，如文本内部特征研究 (词性标注、分词等)，文本分类、聚类，特征提取 (命名实体识别、关键词抽取)，知识图谱应用等。首先他可能会检索“自然语言处理”，得到系列该领域的文献概述，通过大量阅读获取到各个分支领域的信息，张三感觉关键词抽取、命名实体识别两个话题比较感兴趣，他则会依次检索“关键词抽取”、“命名实习识别”获取相关的信息 (如该领域研究现状、出色的学者及相关的文献)。总之，科研人员会通过反复检索关键词获取目标信息，以此开展后续的研究。因此本文将用户搜索的关键词与该关键词匹配的文献关键词

进行关键词网络构建，该网络的功能主要是用来检测用户学者的搜索兴趣和明确检索意图，通过基于图的算法挖掘关键词网络，自动预测他可能会感兴趣的研究内容，进而也为后续给他推荐相似的学者奠定基础。

3.2.1 构建关键词网络 KCG

构建截图，关键词网络的作用，数据节点和边数表

本节主要介绍关键词网络 *KCG* 的构建过程，根据用户搜索的关键词，系统会匹配数据库里的数据，将包含搜索关键词的文献记录返回，通过抽取出文献的关键词，将关键词以共现的关系构建出关键词网络图，即 *KCG* 网络图。例如用户搜索关键词为 "*HFMD*"，返回包含 "*HFMD*" 的文献摘要，动态抽取出每篇文献摘要的关键词，再将这些关键词以共现关系构建成 *KCG* 图（为了加快计算过程，实现快速推进的效果，本文实现先离线抽取出每篇文献摘要关键词，用户搜索后直接返回关键词集合），共现窗口大小 *window* 为摘要的长度。下表**??**为在四个数据集上分别构建的词共现图的节点数目和边数目详细信息，在手足口病数据集上的节点数为 643，边数为 7705，在癌症数据集上的节点数为 164325，边数为 4223234。可见在癌症的 *KCG* 网络图非常大。

3.2.2 挖掘 KCG 网络核心点

3.3 学者 - 关键词模型

本文将学者推荐问题建模在二分图上，该二分图网络的节点包括被推荐的对象，即论文作者 (*Author_node*)，和论文的关键词 (*Keyword_node*)，如果 $Keyword_i$ 出现在作者 $Author_i$ 的论文中，则 $Keyword_i$ 和 $Author_i$ 存在一条边相连，该图是一个无向无权图，为了描述方便，本文将该二分图简称为 AKG。通常情况下，推荐系统是在寻找用户 Users 集合中每个 $user_i$ 和被推荐对象 Authors 集合中的 $author_j$ 之间的关系, 例如可以通过计算的方式给每对 ($user_i$，$author_j$) 组合进行打分，这个分数就可以被用来衡量用户 $user_i$ 可能对被推荐对象 $author_j$ 感兴趣的程度，本文用兴趣即关键词作为联系 $user_i$ 和 $author_j$ 两者的桥梁。本文尝试采用 PersonalRank 算法对构建的 AKG 二分图中的节点进行打分，最后将 Top K 个得分较高的论文

作者作为推荐对象推荐给用户。

3.3.1 构建学者 - 关键词 AKG 网络

通过上段的描述可知该 *AKG* 网络图的构建方式，用邻接矩阵 $M = (M_{i,j})$ 来表示 *AKG* 的权值和边，即 M 定义如下 (**??**), 本文通过模拟用户检索关键词的过程，

3.3.2 Top K 推荐打分算法

为了发现被推荐对象与用户之间关系，目前已经有很多文献提出的算法是基于图结构的数据上，例如 Gori 发表在 IJCAJ 会议上的一篇关于电影推荐的论文[?]，Gori 将电影推荐问题建模在图上，并且改进 PageRank 算法后提出 ItemRank，采用 ItemRank 算法对图中的节点进行打分，本文借鉴其思想，采用 PersonalRank 算法对 AKG 网络图进行打分，最后给用户推荐分值高的前 Top K 个可能感兴趣的论文作者。PersonalRank[?] 算法于 2002 年被 Haveliwala 提出，是一种基于随机游走的图算法，该算法与经典网页重要性排序 PageRank 算法非常相似，为网络中的每个节点计算重要性得分，并且文献[?] 表示其在网络图中表现出很好的性能。PersonalRank 算法的计算公式如下 3.1所示：

$$\mathbf{PR(i)} = \frac{(1-\alpha)}{r_i} + \alpha \sum_{j \in in(i)} \frac{PR(j)}{|out(i)|} \quad r_i = \begin{cases} 1 & i = u \\ 0 & i! = u \end{cases} \tag{3.1}$$

3.4 本章小结

第 4 章　基于聚类的关键词抽取

关键词抽取任务通常包括三个基本步骤，实验的方法多种多样，可基于句子抽取，也可以基于内容理解的抽取，又或者是基于结构的抽取方法等等，无论什么方法被采用，基本上都会面临三个重要的问题，其一是文档冗余信息的识别及处理；其二是对核心关键词的定位识别；其三则是生成的词或短语的可读性和连贯性。宗成庆老师[?] 对识别冗余信息的方法总结为两种.

4.1　数据预处理

对每个数据集，本文都使用以下的数据预处理步骤：首先对文本进行句子切割、分词和词性标注。对候选短语的打分采用如下策略，如公式 (4.1)

$$PhraseScore(p_i) = \sum_{w_j \in p_i} WordScore(w_j) \tag{4.1}$$

4.2　对比算法

TFIDF(term frequency-inverse document frequency) 是基于统计的算法，对论文的摘要进行关键词抽取，TF 是指该词在文本中出现的频率，可以用来描述文档内容；IDF 是该词的逆文档频率，是用来衡量该词区分分档的能力。TFIDF 的为两个公式的乘积，如 (4.2) 所示。

$$TFIDF = \frac{n_i}{N_j} * \log \frac{|D|}{|D_i| + 1} \tag{4.2}$$

其中 N 表示文档 j 中包含的词的总量，n 表示词 i 出现在文档 j 中的词频。$|D|$ 表示语料库中总的文档数量，$|D_i|$ 表示语料库中包含词 i 的文档数量。

TextRank 基于词共现的特征 -TextRank

PositionRank 基于位置的特征 -Position PageRank

TopicRank 本文是基于 TopicRank 的一个改进算法，首先通过将候选集中的

短语通过聚类划分为主题，即主题便是由候选集中的短语的子集表示，再将聚类后得到的主题作为构建完全图的节点，该图的边则是通过计算不同主题之间词与词之间的语义关系得到图的权重，最后采用基于随机游走的 PageRank 算法为每个主题的排序，最后从每个主题中选出一个候选词作为关键词。该特征的获取需要进行的计算包括主题聚类、构建图和 PageRank 给节点打分，因此，相比较前面几种算法而言，该算法的速度相对缓慢。接下来将着重介绍 TopicRank 识别和定义主题的原理及本文对其进行改进的点。

4.3 关键词抽取聚合排序算法

4.3.1 算法描述

4.3.2 算法实现

Borda Count Schulze 方法 Weighted Majority Voting

4.4 本章小结

第 5 章　实验结果与分析

5.1　数据集介绍

本文设计了两组实验，一组是关键词抽取，另外一组是学者推荐。因此，需要对两组实验的实验结果分别进行验证。关键词抽取部分，本文选用 KDD、WWW 以及 Hulth2003 三个数据集上进行实验，在关键词抽取部分的实验评估中，都将作者给定的元关键词作为正确的标准，以上数据的详细内容可查看表5-1。为了验证本文提出的学者推荐模型的可用性，本文通过搜索关键词从微软学术官网上爬取学术论文数据集进行实验[1]。

5.1.1　KDD 和 WWW

数据集 KDD 和 WWW 分别是从知识发现与数据挖掘 (Knowledge Discovery and Data Mining) 顶级会议和万维网 (World Wide Web Conference) 上搜集的真实科技论文数据，这两个数据都包括论文的题目、摘要、作者给定的关键词三个字段。

5.1.2　Hulth2003

数据集 Hulth2003 来自于 Inspec 数据库，Hulth 在 2003 年时搜集的从 1998 年到 2002 年期间的期刊论文，该数据集中包含摘要和关键词，每篇摘要都被人工分配了两个关键词，其中一个集合中的关键词是 Inspec 数据库的词库中能已经包含的词，但另外一个集合是专家认为跟这篇论文比较匹配的关键词就能被确定为关键词，而不会受到 Inspec 词库的限制，本文的实验选用后一种关键词集进行实验。

表 5-1　数据集

数据集	类型	语言	数量
WWW	论文摘要	英文	1330
KDD	论文摘要	英文	755
Hulth03	论文摘要	英文	500

[1]https://academic.microsoft.com

5.1.3 微软学术论文数据

本文通过搜索“hfmd”(手足口病)、“t2dm”(二型糖尿病)、“pneumonia”(肺炎) 和“cancer”(癌症) 四个关键词从微软学术网站上爬取了 2015 年到 2017 年的生物医学文献数据。其中包括论文标题、摘要、作者三个属性。

5.2 评价指标

5.2.1 关键词抽取评价指标

在关键词抽取工作中，提出过很多评价方法，其中最为常用的是准确率 (Precision)、召回率 (Recall)、准确率和召回率的调和平均数 F1 值 (F1-score 和 MRR，本文也选用这三种评价指标验证关键词抽取算法的有效性。

准确率又名查准率，本文中通过将算法自动抽取的关键词和人工标记的关键词进行交集运算，例如：同一篇论文摘要，算法自动抽取出的 Top 4 关键词集合为 $E(k)$ = {”$hfmd$”,”$EV71$”,”$a16$”,”$encephalitis$”}，人工标注的关键词集合为 $S(k)$ = {”$hfmd$”,”$enterovirus71$”,”$coxsackievirus$”,”$Picornaviridae$”}，两个集合的交集为 {“hfmd”}，所以正确抽取的关键词个数为 1 个。因此，准确率、召回率和 F1 值的公式 (5.1) (5.2) (5.3) 所示：

$$Precision = \frac{E(k) \cap S(k)}{E(k)} \tag{5.1}$$

$$Recall = \frac{E(k) \cap S(k)}{S(k)} \tag{5.2}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.3}$$

5.2.2 学者推荐评价指标

在学者推荐实验中，我们在微软学术数据上进行实验，针对无标签数据，本文通过采用对比的方式进行结果评价，即将本模型的推荐结果与微软学术的推荐

列表进行对比。该设计的灵感源自于关键词抽取的评价标准，将微软学术官网的学者推荐列表作为正确结果，当用户搜索关键词后，微软学术会根据搜索的关键词推荐相应的学者，因此本文在无标签的情况下，为了验证模型有效性，就通过与权威的结果进行比对以此证明实验的可行性。微软学术的推荐结果可能是综合考虑多项指标之后，例如文献的引用次数、下载量、影响因子，作者的影响力、发文数量等，本文是在仅仅只有文本数据，而没有其他相关指标的情况下设计的推荐模型。通过查阅文献还发现，部分学者也曾在文献中提出将自己算法的实验结果截图和比较权威的机构或者公司的结果截图作定性的对比，针对无标签的数据，该设计方案有一定的迁移意义。下图 5-1左侧即为在微软学术官网搜索关键词”t2dm”(二型糖尿病) 后推荐 20 个学者的列表，右侧为检索”t2dm” 时相关的文献列表。本文取 Top20、Top10、Top5 三种结果，若本模型推荐的前 10 个学者中有 6 个在微软学术推荐的列表中出现，那么该模型的正确率为 P(top10) = 6/10，召回率为 R(top10) = 6/20。同理，若本模型推荐的前 5 个学者中有 2 个在微软学术推荐的列表中出现，那么该模型的正确率为 P(top5) = 2/5, 召回率为 R(top5) = 2/20。因此，学者推荐的评价公式准确率、召回率和 F1 值的公式如 (5.4) (5.5) (5.6) 所示：

$$P = \frac{R(u) \cap T(u)}{R(u)} \tag{5.4}$$

$$R = \frac{R(u) \cap T(u)}{T(u)} \tag{5.5}$$

$$F1 - score = \frac{2 * P * R}{P + R} \tag{5.6}$$

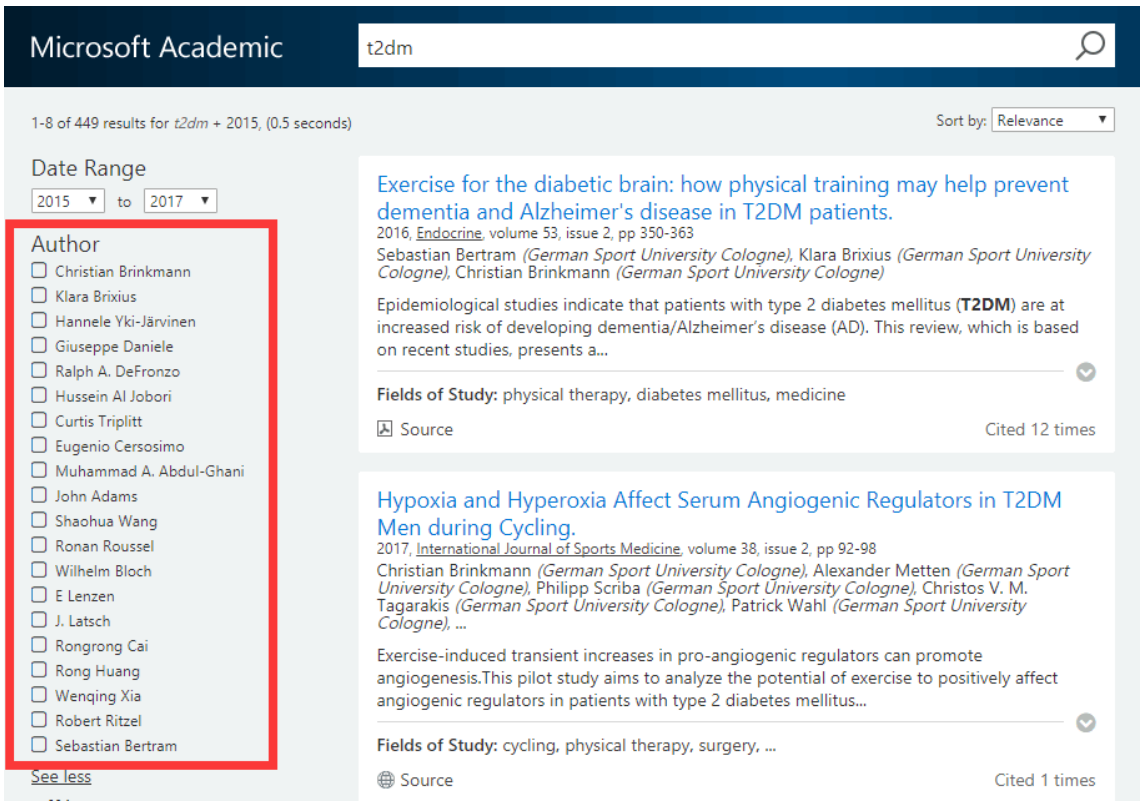

图 5-1　红色方框列表则为在微软学术官网检索关键词“t2dm”时的推荐结果

表 5-2　关键词抽取准确率结果

Dataset	Method	Top2	Top4	Top6	Top8
KDD	TF-IDF	11.4	8.3	6.5	5.5
	TextRank	9.5	8.1	7.1	6.3
	融合后的算法	0.55	0.2	0.15	
WWW	TF-IDF	12.2	9.1	7.0	5.8
	TextRank	11.7	10.1	8.4	7.4
	融合后的算法	0.55	0.2	0.15	
Hulth03	TF-IDF	0.5	0.15	0.1	
	TextRank	0.55	0.15	0.0	
	融合后的算法	0.55	0.2	0.15	

表 5-3　关键词抽取召回率结果

Dataset	Method	Top2	Top4	Top6	Top8
KDD	TF-IDF	5.6	7.9	9.2	10.2
	TextRank	4.6	7.7	9.9	11.8
	融合后的算法	0.55	0.2	0.15	
WWW	TF-IDF	5.4	7.3	8.4	9.2
	TextRank	4.8	8.2	10.1	11.7
	融合后的算法	0.55	0.2	0.15	
Hulth03	TF-IDF	5.6	7.9	9.2	10.2
	TextRank	0.55	0.15	0.0	
	融合后的算法	0.55	0.2	0.15	

表 5-4　关键词抽取 F1-score 结果

Dataset	Method	Top2	Top4	Top6	Top8
KDD	TF-IDF	7.5	8.1	7.6	7.1
	TextRank	6.2	7.9	8.3	8.2
	融合后的算法	0.55	0.2	0.15	
WWW	TF-IDF	7.1	8.1	7.7	7.1
	TextRank	6.8	9.0	9.1	9.1
	融合后的算法	0.55	0.2	0.15	
Hulth03	TF-IDF	0.5	0.15	0.1	
	TextRank	0.55	0.15	0.0	
	融合后的算法	0.55	0.2	0.15	

5.3　实验结果与分析

5.3.1　关键词抽取结果

5.3.2　学者推荐结果

5.3.3　结果分析

5.4　本章小结

表 5-5　使用多种算法从关键词网络中选出核心节点的推荐结果表。

Dataset	Method	Top20	Top10	Top5
HFMD	Degree Centrality	0.5	0.15	0.1
	Closeness Centrality	0.55	0.15	0.0
	Eigenvecter Centrality	0.55	0.2	0.15
	Jaccard	0.4	0.35	**0.25**
	Wang's method	**0.55**	**0.35**	0.2
	Liu's method	0.45	0.25	0.0
T2DM	Degree Centrality	0.2	0.2	0.2
	Closeness Centrality	0.25	0.2	0.2
	Eigenvecter Centrality	0.2	0.0	0.0
	Jaccard	0.2	0.15	0.0
	Wang's method	0.2	0.0	0.0
	Liu's method	**0.35**	**0.35**	**0.2**
Pneumonia	Degree Centrality	0.15	0.1	0.05
	Closeness Centrality	0.15	0.1	0.05
	Eigenvecter Centrality	0.15	0.1	0.05
	Jaccard	**0.15**	**0.15**	**0.05**
	Wang's method	0.15	0.1	0.05
	Liu's method	0.15	0.1	0.05
Cancer	Degree Centrality	0.05	0.0	0.0
	Closeness Centrality	0.05	0.0	0.0
	Eigenvecter Centrality	0.05	0.05	0.05
	Jaccard	**0.15**	**0.1**	0.0
	Wang's method	0.05	0.05	**0.05**
	Liu's method	–	–	–

第 6 章　总结与展望

数据集不充分，数据获取不全，可获取数据有限，使得模型应用中部分特征维度得不到数据支持

在未来笔者将对该研究课题进行更为深入的研究与探索第一，进一步完善推荐模型中的组成模块，挖掘更多可以表达学者需求的内容特征；第二，笔者将尝试采用更为充分的数据集与数据量，对模型的实现进行更精确得仿真模拟，并且采用更有效的数据清晰、预处理方法，减少因数据、参数等降低推荐准确度的影响。第三，采用深度学习相关技术提取文本特征

表 目 录

图 目 录

致　谢

光阴似箭，日月如梭，三年的时间，我的硕士生涯已接近尾声，回头想想这段短暂的求学路，时而喜悦，时而惆怅，感谢命运的安排，让我有幸结识了许多良师益友。首先感谢我的恩师朱佳教授，谢谢您一直以来对我的照顾，让我有幸遇到您这样对学生亦师亦友的好导师，您和蔼可亲、学识渊博、没有架子以及乐观向上的工作生活态度深深感染着我。感谢您耐心指导我的论文工作，感谢您陪我走过人生最重要的结婚典礼，感谢您带领我们参加各种会议开阔视野，感谢您带我们吃各种美食，感谢您跟我们聊各种八卦新闻。。。再次向您致以衷心的感谢和崇高的敬意。

感谢学者网团队的各位小伙伴，肖丹阳、林雪琴、韦经敏，感谢你们的 3 年陪伴和对我的学术帮助，感谢 332 实验室的师弟师妹们，杨芬、汪序明、郑泽涛、伦家琪、胡迎彬、于晗宇，感谢你们带给我的欢乐和帮助，感谢已经毕业的师兄师姐，董浩业、丁蕊、陈凌潇、许传华、孔剑龙，跟你们在一起真的很开心快乐，感谢生活学习中有你们。

感谢我的最佳室友们男神玉、波波、嘉良，跟你们在一起的日子真的很幸福，记得大半夜去外面吃烧烤喝啤酒，记得去自助 KTV 嗨翻全场，记得在宿舍吃着零食聊学术八卦到深夜，这些美好的时光将是我一生宝贵的财富，谢谢你们的陪伴，让我的研究生生活丰富多彩。

感谢我的父母、岳父母、爱人、女儿及所有的亲朋好友，你们是我永远的支持者，让我在学习和生活中都受到你们的无私关怀和帮助，才能完成研究生求学生涯，感谢你们的付出，让我开心的度过每一天。

感谢所有的 2016 级的同窗们，谢谢你们陪我一起成长，祝各位同窗前程似锦；感谢我的母校华南师范大学，谢谢您为我们提供美好的校园环境和学习氛围，祝您的办学越来越好；最后感谢各位答辩评委老师，谢谢你们为我指点出不足，祝各位老师永远健康快乐。

作者攻读学位期间发表的学术论文目录

发表的学术论文

[1] Hailin F, Jianguo L, Jiemin C, et al. Sequence-based Recommendation with Bidirectional LSTM Network[C]. PCM 2018: 428–438.(第一作者)

[2] Jiemin C, Jianguo L, Jing X, et al. A Hybrid Collaborative Filtering Model: RSVD Meets Weighted-Network Based Inference[J]. 網際網路技術學刊 2016: 1221-1233.(合作作者)