中图分类号：O157　　单位代号：10280
密　　级：公开　　学　　号：18724781

硕士学位论文

SHANGHAI UNIVERSITY MASTER'S DISSERTATION

题目	社交网络数据采集与分析

作　　者　王鹏
学科专业　计算机视觉
导　　师　王宜敏
完成日期　二〇一八年十月

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名:

主席:

委员:

导　　师:

答辩日期:

原创性声明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名: _____ 日期: _____

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定。即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

签名: _____ 导师签名: _____ 日期: _____

上海大学工学硕士学位论文

社交网络数据采集与分析

作　　者：王鹏
导　　师：王宜敏
学科专业：计算机视觉

计算机科学与技术学院
上海大学
2018 年 10 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in Science

An Introduction to LaTeX Thesis Template of Shanghai University v2.0

Candidate: ahhylau

Supervisor:

Major: Operational Research and Control Theory

Department of Mathematics

Shanghai University

May, 2017

摘 要

近些年来，随着社交网络服务蓬勃发展，互联网世界快速进化，社交网络在全球拥有着大量用户，并且使用人数在日益增加，已然成为影响巨大的信息平台。掌握其中有利的信息例如社交网络中用户的行为特征、所具有以及发布的信息所具有的传播规律，甚至内容的定位等，不仅能帮助企业根据客户所需制定出更好更完善及用户友好型的产品，提供更有力的服务，还可以进行更加有效的网络媒体营销。另外，还可为政府及相关部门在舆论控制方面提供有利理论依据和执行的条件，从而优化社会风气。

所以，我选择了基于社交网络的数据采集以及分析。本文中提出并实现了一种，基于新浪微博程序接口和网络数据流相互结合的方式进行数据的采集文献，利用cookie对新浪微博网页端进行了模拟登陆，实现了数据抓取采集。并且通过监测信息交互时传输的数据包。进行动态抓取页面信息。利用python语言网络爬虫和json等数据库函数，解析HTML XML，获得用户的数据信息。然后通过基于图的聚类的社交网络算法以及PageRank算法，进行用户节点数据的建模。收集用户的偏好以及计算其节点的重要性，对微博中用户与用户之间的社交网络关系进行分析，通过多种不同算法对节点的重要性进行了分析，对比多种算法之间的精准度，对其间的影响力评价模型进行了实现。

本文中所讨论的社交网络用户数据信息的获取，主要问题是在于解决了微博近些年来日渐难以爬取大量数据的问题并通过数据挖掘对对用户行为进行分析研究，得到了具体的用户影响力公式，并根据实际进行了验证与误差分析。

关键词: 社交网络数据采集数据处理与分析数据挖掘图聚类

ABSTRACT

Abstract in English.

Keywords: TEX; LATEX; Template; Thesis

目 录

主要符号对照表

$\mathcal{T}$	张量
H	超图
$\mathcal{A}(H)$	超图 H 的邻接张量
$\mathcal{L}(H)$	超图 H 的拉普拉斯张量
$\mathcal{Q}(H)$	超图 H 的无符号拉普拉斯张量
ρ	谱半径
G	图
κ	连通度
χ	染色数
Δ	最大度
δ	最小度

第一章　绪论

1.1　研究背景

社交网络源自网络社交，其起始为电子邮件，互联网的本质上就是计算机之间相互的联网，早期的e-mail仅解决了远程的邮件传输问题，并流传使用至今，同时它也是网络社交的起点。其进一步的发展出的成果BBS，进一步的实现了人们消息之间进行的互通与交流。BBS把网络社交推进了一步使得即时通信和博客成为了社交工具的升级版。他们提高了信息传输速度和并行处理能力。解决了信息发布单一的特点。伴随着网络社交的演变进一步进化，社交网络随之出现。

1.2　研究意义

如果能够有效地分析，社交网络中用户行为、用户特征、用户之间的信息相互交流，并掌握其行为模式以及交流方式，不仅能够帮助运营商更加全面掌握用户需求，从而更新产品、提高用户体验，还能够帮助卖方更好的了解买方需求，从而采取更加有效的网络传销方式和手段，更加吸引买家的注意力，提高用户使用感。进而有助于推动经济发展。并且，社交网络数据采集与分析能够帮助政府及其有关部门在网络媒体上进行合理的舆论宣传，并且可以对网络安全方面进行监控，进而实现绿色上网。

1.3　研究现状

目前，国内对社交网络的研究主要是从社交网络的分析的方向出发，其数据的抓取方法主要包括，搜索引擎爬虫抓取数据，http响应时间所获取的数据，或者网络流量监测来源数据。例如近些年来国家科技研究部门的最新成果BSNiner，其主要功能是可以通过同时应用多种的计算方式从而获取大量且种类各异的数据的处理结果，并且精确度极高；青大企业的自主研发产品iCNiner，其运行速度和精准度已经达到了国际要求的标准。此外我国政府还加大对数据挖掘等相关方面的研发以及投资力度和人才培养，并在全国多所高等院校内成立相关研究机构。

目前,国外数据挖掘方面的最新研究发展主要有对发现相关知识的方法的进一步研究,例如近年来注重对Bayes(贝叶斯)方法以及Boosting方法方面的研究，改进提高;将KDD与数据库进行紧密结合;利用传统的统计学回归方法在KDD中进行应用。在应用方面主要体现在利用KDD等相关商业软件工具从解决孤立的问题过程转向建

立问题解决的整体系统,其主要用户包括保险公司、大型银行和销售业等相关企业。许多计算机公司和研究机构都非常重视对数据挖掘相关的开发应用，譬如IBM和微软都相继成立了相应的研究中心。

第二章　数据挖掘相关概念

2.0.1　数据挖掘的基本概念和原理

对于数据挖掘可以从两个方面进行看待和定义。首先是技术方面的定义，数据挖掘是通过对大量数据进行分析从而发现并提取隐含在其中的具有价值的信息的过程。其次对于商业角度来说，数据挖掘是一种新型的商业信息处理技术，可以从海量数据中进行数据抽样并进行分析以及模型化处理从而提前有利于商业发展的关键数据。通常数据挖掘的任务主要分为两大类：预测任务，描述任务。

数据集的三个主要特性也是研究的重要方面对数据挖掘技术具有重要影响。其分别为维度，稀疏性和分辨率。

- 维度：数据集的维度是数据的集中的对象所具有的属性个数。分析较高维度的数据时有可能会出现维灾难。因此，数据预处理的一个重要目的即为降低数据维度。
- 稀疏性：指所对应对象的大部分属性上的值为0。稀疏性有利于节约计算和存储的时间。
- 分辨率：即衡量数据的尺度，在不同的分辨率下对相同的数据进行分析，得到的性质可能是不同的。
- 数据挖掘是多学科交融所产生的领域，其不仅利用了来自统计学的抽样，估计和假设检验。并且还运用了机器学习，人工智能，以及模式识别的数据搜索算法，建模技术，并且还和分布式计算，可视化处理有密切关联

2.0.2　数据挖掘常用方法与功能

数据挖掘主要有四种方法与任务：

1. 预测建模：为了说明所处理函数为所预测的目标函数变量而进行的建模。主要分为，分类以及回归。两者都是主要是为找出具有相同特性的一部分数据集。前者主要针对离散型数据，后者则主要针对连续型数据。
2. 关联分析：用来分析并找出一批数据集中所隐藏的可能所具有的强关联关系。并以特征子集或者是蕴含关系的形式表示出来。
3. 聚类分析：其主要目的在于发现一组具有相似特性的数据集并将其聚集在一起。聚类可用于对社交网络中用户的行为社团进行分组。
4. 异常检测：识别具有明显不一致特征的数据集。其基本方法是找出实际结果与预测建模所得出结果中相差较大的数据。

第三章　基于用户之间的社交网络分析的算法

3.1　基于PageRank算法的用户影响力算法

3.1.1　PageRank算法原理

PageRank算法起初是Google搜索引擎中用来对计算机网络中所涉及到的网页所具有的影响力进行排序的算法，运用该算法提高了Google搜索的搜索结果的质量以及搜索信息的准确性。

该算法默认给每个网页一个初始的默认的PR值，当一个网页被其他网络所链接时，其所对应的PR值会增大，其影响力排名会更加靠前。并且影响力靠前的网页对其他网页对PR值的改变具有更大的作用，即当一个网页被排名高的网页所链接时，其排名的上升将会更加快速。根据这种思想，PageRank算法所建立的数学模型如下所示：

$$PR_i = \sum_{j,i \in E} \frac{PR_i}{O_j} \tag{3.1}$$

即一个网页的PR值等于所有连接到该网页的PR值的加权平均数。其中O_j表示网页j的向外链接数,如图3.1:

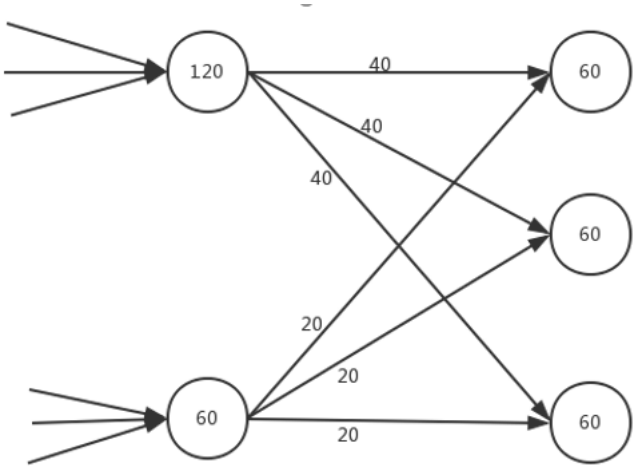

图 3.1　PageRank算法抽象图

3.2　算法实现的流程与描述

(1) 从爬取所得到的数据集之中读取社交网络中各用户数据，并对各用户的PR值进行初始化设置，同时设置误差允许的范围p

(2) 根据用户之间的关系，利用PageRank算法对各用户之间的关系以及影响力进行计算。并且将用户自身的影响力有效分配给关注的用户。
(3) 再次对用户的影响力进行计算，直到与上轮误差达到预先设置的p值之内。这时可以认为读出来数据是准确的，并可以将其存入数据库以便进行下一步的继续研究。

3.3 基于图聚类的用户影响力算法

3.3.1 基于图聚类的算法

图聚类算法原始是从聚类算法中延伸出来的，其中的数据对象通过结点来表示，并且用对应结点之间边的权值表示两个数据对象之间的邻近度。图聚类算法兼顾了聚类算法的核心思想并且利用了图的许多重要特性以及性质。下面是一些图聚类所用的重要方法。

(1) 稀疏化邻近度图，只保留对象与其最向临近对象的链接。这种的稀疏化主要用于处理噪声和离群点。并且通过稀疏化也可以使得对稀疏图进行有效图划分算法。
(2) 基于共享的最近临个数。定义两个对象间的相似性的度量。这种方法主要用于在克服高维和变密度簇方面。
(3) 定义核心对象并构建环绕其间的簇。通过引进邻近度图或者稀疏化的邻近图的基于密度的分析，围绕核心对象并构建簇将会进一步发现不同形状和大小的簇。
(4) 使用邻近度图中所显示的信息，从而对两个簇是否应当合并作出更复杂的评估。

3.3.2 基于图聚类的社交网络聚类算法思想

微博属于一种新兴的社交网络，拥有着成簇特性。本实验采用了社团发现算法，该算法类似于群的概念，团簇内部用户通过爱好，实际中的社交关系等联系在一起。且团簇彼此之间受影响较小。因此可以将各个团簇从微博网络中独立出来进行研究，对其利用复杂网络的聚类算法对社交网络用户进行聚类，分解出其团聚结构。

本次实验所用的是一种改进的进行了平衡聚类事件和聚类精度的GN算法对社交网络用户行为进行聚类。该改进算法中将链接聚类系数定义为包含该路径的所有短回路数，计算i结点到j结点边的连接聚类系数公式如式3.2：

$$C_{i,j}^{g} = \frac{Z_{i,j}^{g} + 1}{S_{i,j}^{g}} \tag{3.2}$$

并且通过模块度函数用于评价网络聚类的终止条件，式3.3:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \sum_i a_i^2 \tag{3.3}$$

本文使用的是Radicchi所提出的通过改进后的GN 算法对社交网络进行聚类，并且使用模块度函数作为聚类终止的判断条件。其算法流程如下：

(1) 首先通过微博关注者与被关注者之间关系构建微博网络

(2) 计算Q值并通过判断是否终止聚类来进行划分

(3) 计算网络中所有连接边的连接聚类系数

(4) 删除连接聚类中系数最小的边

(5) 重复步骤(2)

在对所搜集到的万级数据中对数据进行聚类，对Q取值为0.6。其所聚类迭代的次数以及所获得的团簇个数如表3.1所示。

表 3.1　迭代次数与团簇个数

迭代次数	团簇个数
0	1
2	32
4	471
6	724
8	1534

从团簇的个数变化中可以看出，前6次的迭代中团簇分裂速度很快，第七次第八次分裂速度逐渐变慢最终分裂为1543个团簇。对其变化速率进行分析，发现其原因是在第七次迭代开始，社交网络中已经出现了大量满足Q值的团簇停止了其分裂过程。团簇内部节点个数分布情况如表3.2所示

表 3.2　团簇内部结点个数

簇内结点数	簇数
100及以下	1382
100-200	102
200-300	33
300-400	6
400以上	1

其中大部分的团簇中所包含的微博用户都少于100个，并且随着结点数的增加，所符合的团簇数目逐渐变小。团簇中最大结点数为478，最小值为1.其中90%的团簇

结点数少于100个。

3.4　基于用户的微博团簇影响力计算

本实验中主要所采取进行横向对比的是聚类的度中心度分布，介数中心度分布，距离中心度分布，距离中心度分布以及通过pagerank算法进行计算得出的网络用户影响力分析。

(1) 粉丝数量：用户偏好在此次中即为用户对其他微博用户的关注度，对微博内容信息的转发评论等表示偏好的行为，当一个用户的粉丝数目越多时，其所发布的微博内容将会对更多的用户者偏好产生影响。因此大部分微博用户影响力评价模型中也将此作为一个计算因子。

(2) 微博数量：在社交网络中，用户发布微博是产生自身印象力并能对其他用户偏好产生影响的最好办法。微博的发布可以作为用户偏好以及影响力的传播载体，并且也可以体现用户在社交网络中的活跃程度。因此也将微博数量作为影响力计算因子。

本实验中主要所采取进行横向对比的是聚类的度中心度分布，介数中心度分布，距离中心度分布，距离中心度分布以及通过pagerank算法进行计算得出的网络用户影响力分析。下面用图3.2表示说明节点的各个性能：

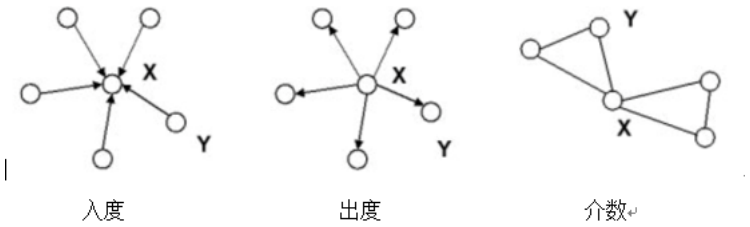

图 3.2　中心度的不同表示观点

(1) 度中心度：是在网络分析之中刻画节点中心性的一个直接度量指标，一个节点的度中心度越大，则其中心性越高，在社交网络中所处的地位也就越重要。

(2) 距离中心度：决定了一个节点的紧密性。即该节点到达其他节点的难易程度。该性质是基于某一节点与社交网络中所有其他节点之间的平均最短路径长度计算而得到的。

(3) 介数中心度：核心思想是两个非邻接成员间的相互作用依赖于社交网络中的其他成员，特别是依赖于两成员间路径上的成员，它们对两个非邻接成员之间起着某种控制或依赖关系。如果一个成员A位于其他成员的多条最短路径上，那么成员A的作用就比较大，也具有较大的介数中心性。其本质为网络中包含成

员B的所有最短路径条数占所有最短路径条数的百分比。

$$C_B(i) = \frac{\sum_{j<k} g_{jk}(i)}{g_{jk}} / [\frac{(n-1)(n-2)}{2}] \tag{3.4}$$

其中g_{jk}表示连接j-k的最短路径的条数，$g_{jk}(i)$表示位于最短路径的个数。

第四章　模型实现以及结果分析

本节通过构造一个社交网络用户影响力分析模型来实现本次的课程设计。将在算法原理的基础上详细介绍本次课设包含的处理模块，最终将所建立模型的与实际情况作比较进行误差分析。

4.1　抓取到的部分微博数据

	A	B	C	D	E	F	G	H	I	
1	粉丝	字段1_链接	字段2	字段3_文本	字段3_链接	字段4	字段5_文本	字段5_链接	字段6	字段7_文
2	tetsuJP	https://weibo.com/u/304	关注 255	255	https://weibo.com/30426	粉丝74	74	https://weibo.com/30426	微博299	299
3	用户6541408673	https://weibo.com/u/654	关注 100	100	https://weibo.com/65414	粉丝1	1	https://weibo.com/65414	微博0	0
4	用户6527055609	https://weibo.com/u/652	关注 58	58	https://weibo.com/65270	粉丝4	4	https://weibo.com/65270	微博0	0
5	用户6504010375	https://weibo.com/u/650	关注 23	23	https://weibo.com/65040	粉丝1	1	https://weibo.com/65040	微博0	0
6	船载千斤掌舵一人	https://weibo.com/u/641	关注 26	26	https://weibo.com/64103	粉丝8	8	https://weibo.com/64103	微博7	7
7	用户6287017916	https://weibo.com/u/628	关注 164	164	https://weibo.com/62870	粉丝2	2	https://weibo.com/62870	微博0	0
8	A小阿静A	https://weibo.com/u/579	关注 89	89	https://weibo.com/57980	粉丝31	31	https://weibo.com/57980	微博19	19
9	如果遇到彼此	https://weibo.com/u/610	关注 432	432	https://weibo.com/61075	粉丝61	61	https://weibo.com/61075	微博6	6
10	流浪狗0309	https://weibo.com/u/654	关注 24	24	https://weibo.com/65410	粉丝1	1	https://weibo.com/65410	微博2	2
11	逃向北海遛的猫z	https://weibo.com/u/245	关注 279	279	https://weibo.com/24554	粉丝50	50	https://weibo.com/24554	微博3	3
12	M1emory	https://weibo.com/u/378	关注 215	215	https://weibo.com/37844	粉丝383	383	https://weibo.com/37844	微博0	0
13	IIIII好孩子86	https://weibo.com/u/562	关注 45	45	https://weibo.com/56287	粉丝13	13	https://weibo.com/56287	微博5	5
14	毛妞儿999	https://weibo.com/u/654	关注 0	0	https://weibo.com/65414	粉丝1	1	https://weibo.com/65414	微博0	0
15	绿筱媚青莲1	https://weibo.com/u/388	关注 228	228	https://weibo.com/38842	粉丝32	32	https://weibo.com/38842	微博58	58
16	用户6102611353	https://weibo.com/u/610	关注 40	40	https://weibo.com/61026	粉丝4	4	https://weibo.com/61026	微博4	4
17	Jenny_茵	https://weibo.com/u/599	关注 108	108	https://weibo.com/59914	粉丝11	11	https://weibo.com/59914	微博81	81
18	Swider--	https://weibo.com/u/502	关注 435	435	https://weibo.com/50210	粉丝39	39	https://weibo.com/50210	微博112	112
19	用户6541406893	https://weibo.com/u/654	关注 67	67	https://weibo.com/65414	粉丝1	1	https://weibo.com/65414	微博0	0
20	阳光小朋友	https://weibo.com/u/600	关注 28	28	https://weibo.com/60041	粉丝12	12	https://weibo.com/60041	微博27	27
21	用户6541407314	https://weibo.com/u/654	关注 27	27	https://weibo.com/65414	粉丝1	1	https://weibo.com/65414	微博0	0
22	小妍微博	https://weibo.com/u/273	关注 104	104	https://weibo.com/27301	粉丝44	44	https://weibo.com/27301	微博47	47
23	运力乖乖儿	https://weibo.com/u/257	关注 334	334	https://weibo.com/25749	粉丝96	96	https://weibo.com/25749	微博9	9

图 4.1　部分微博数据

4.2　社交网络用户行为分析系统IPO图

IPO图为输入处理加工图的建成，用来说明输入数据，对输入数据的数据加工过程，以及输出数据。本次实验所涉及到的IPO图如图4.2所示

(1) 数据预处理模块

(2) 对复杂网络进行聚类分析模块

(3) 对数据进行与计算模块

(4) 团簇影响力的计算

(5) 微博影响力的计算。输出数据为聚类中度的出度，入度分布，度中心度分布，距离中心分布。

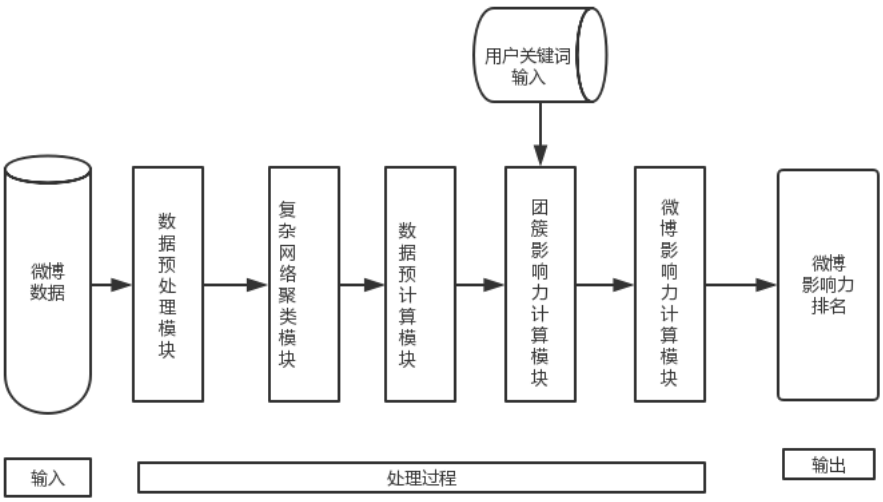

图 4.2　社交网络用户行为分析系统

4.3　社交网络影响力分析系统的实现

4.3.1　数据预处理模块

该模块的主要功能是对爬取到的微博数据进行预处理，主要可分为两部分，如图4.3：

(1) 剔除粉丝数目为零的用户

(2) 剔除僵尸粉

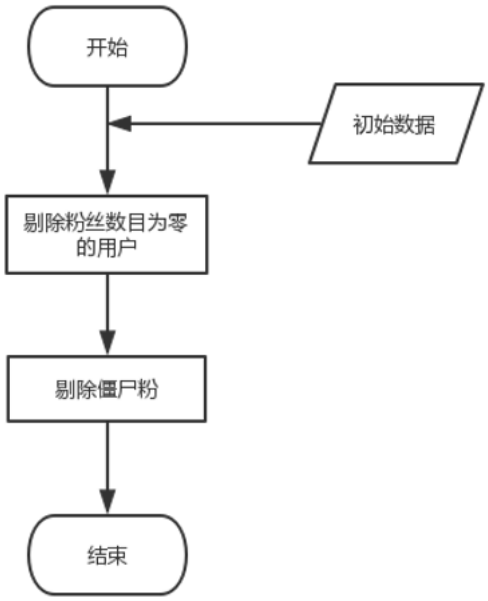

图 4.3　数据预处理流程图

4.3.2　社交网络聚类模块

社交网络聚类模块对上一模块的输出数据进行聚类处理。该部分通过微博连接关系构建社交网络，并且使用改进过的GN算法把社交网络分解为多个相互联系紧密且可以独立研究的团簇。模块的具体计算流如图4.4：数据经过该模块处理完成之

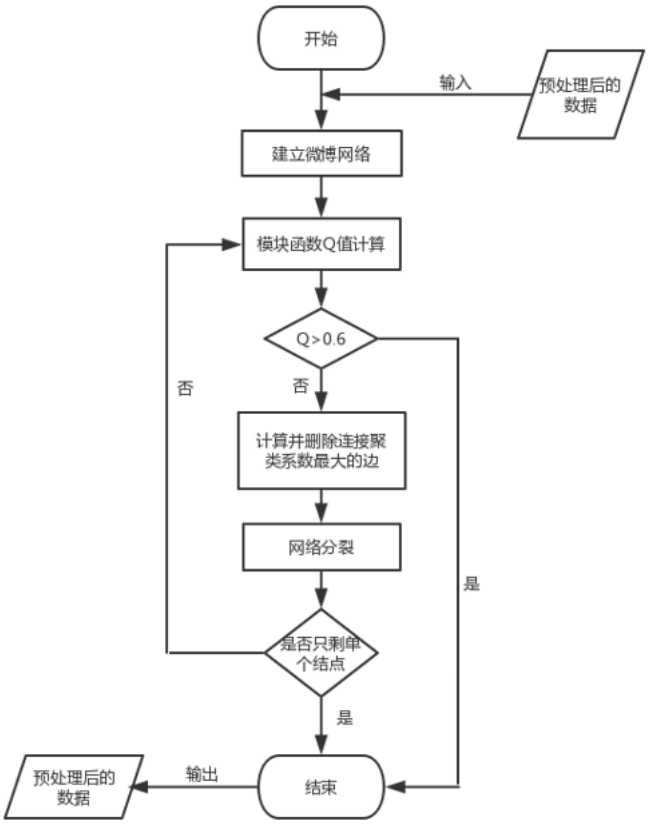

图 4.4　复杂网络聚类模块流程图

后，所有的社交网络用户都将拥有一个标示用于指示用户所属的团簇。

4.3.3　数据预计算模块

在实际计算微博影响力的时候，采用的是用迭代的方法并结合pagerank算法，导致计算较为缓慢。但是其中涉及到的部分数据，例如用户的平均粉丝数，平均微博数等，可以先将这几个参数进行预计算并将结果存入数据库之中，便可以提高系统的计算实时性。

4.3.4　团簇影响力计算模块

此部分的计算模块是对上一步已经进行完毕聚类的团簇进行影响力计算，以此

来作为微博影响力计算的参数。如图4.5

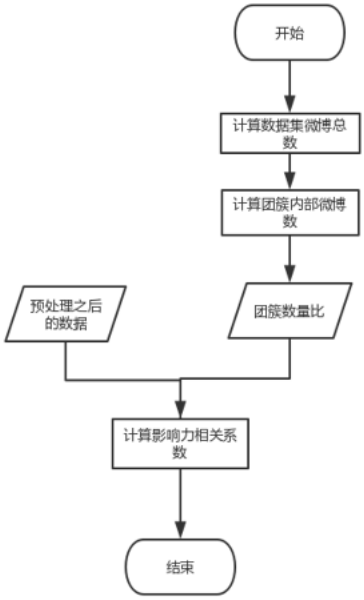

图 4.5　团簇影响力计算模块流程图

4.3.5　社交网络用户影响力计算模块

微博用户影响力计算模块是在前几个已经完成的模块基础上对最终的社交网络影响力进行计算，由于PR值以及团簇内部已经被计算。故在此模块计算之后即可做出最终的社交用户影响力数据相关度分布。其计算模块流如图4.6

4.4　实验结果及分析

对模型的评价在于对利用多种方式对微博偏好影响力计算的实时性进行评估，对其进行横向对比。图4.7-图4.12分别为网络节点中的度性能坐标分析图。

从这些图可以看出，节点的频次数目随着出入度的增加而减少，节点出入度越大，其所对应的中心距离度越大。且随着节点出入度的增加，各类数值频次趋于稳定，在数据集增加到为10$\hat{2}$为级别时，所得出结论逐渐稳定。其中pagerank算法由于其参数方面对用户关键词的计算度量较少因此结果比较不稳定。

在社交网络中，当一个用户节点出入度以及距离中心度等越大时，其用户影响力越大。且在社交网络总体之中此类用户节点所占比例较小。

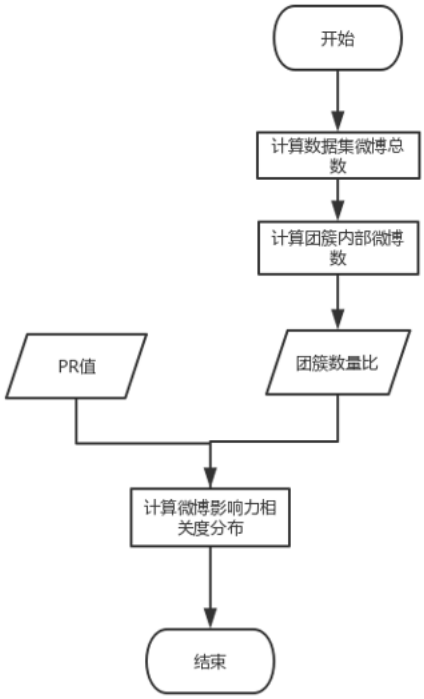

图 4.6　微博影响力计算模块流程图

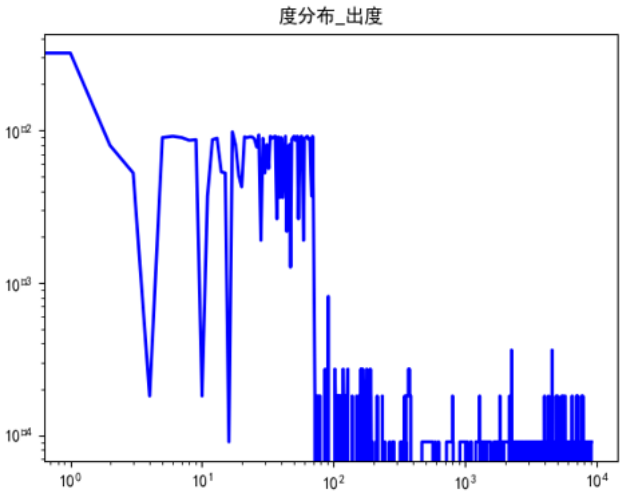

图 4.7　度分布-出度

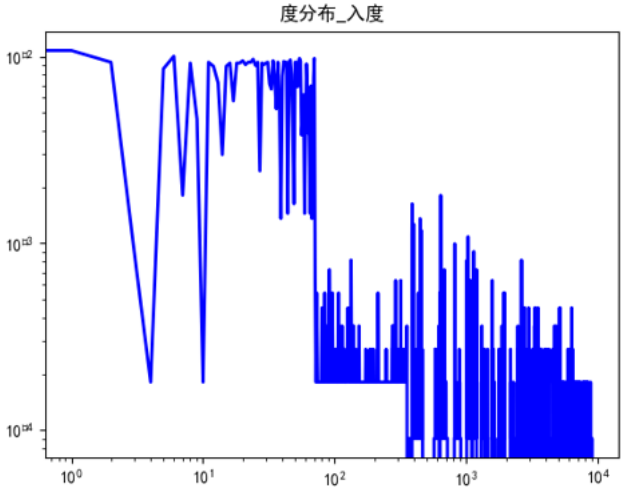

图 4.8　度分布-入度

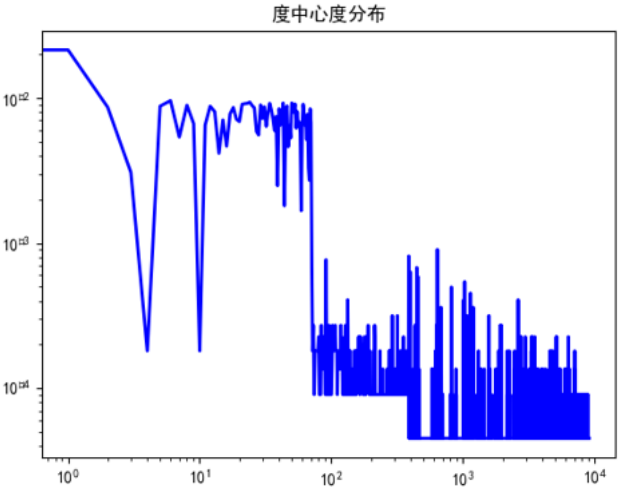

图 4.9　度中心度分布

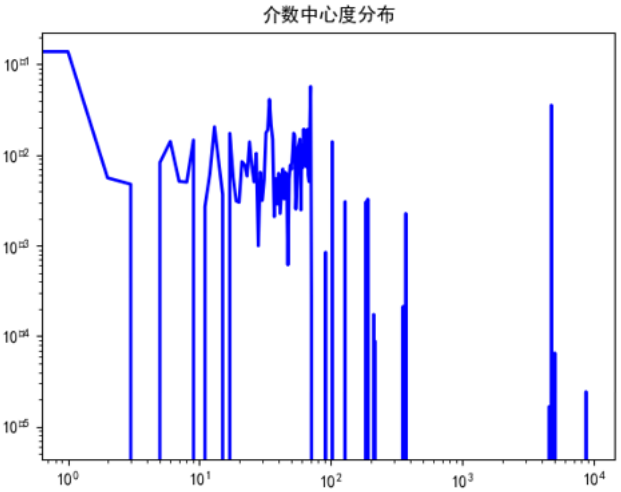

图 4.10　介数中心度分布

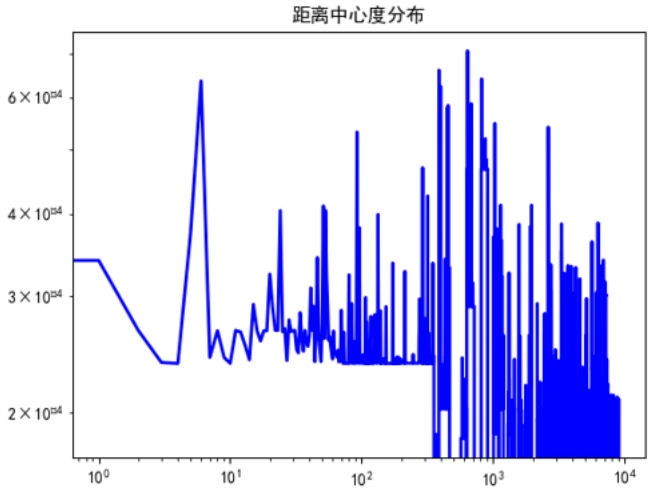

图 4.11　距离中心度分布

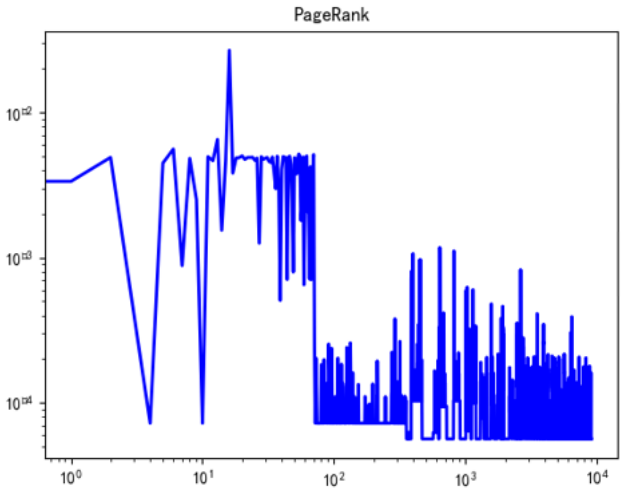

图 4.12　page rank分布

插图索引

表格索引

参考文献

[1] Microsoft Corporation, “FAT32 File System Specification”, http://microsoft.com/whdc/ system/platform/firmware/fatgen.mspx, 2000

[2] Microsoft Corporation, "Extended FAT File System", http://msdn2.microsoft.com/en-us/ library/aa914353.aspx, 2007

[3] M. S. Kwon, S. H. Bae, S. S. Jung, D. Y. Seo, and C. K. Kim, “KFAT: Log-based Transactional FAT File system for Embedded Mobile Systems”, In Proceedings of 2005 US-Korea Conference, ZCTS-142, 2005

[4] Microsoft MS-DOS Programmer’s Reference: version 5.0.”, Microsoft press. 1991.

[5] Liang Alei, Liu Kejia, Li Xiaoyong , “FATTY : A reliable FAT File System”, Proceedings of the 10th Euromicro Conference on Digital System Design Architectures, Methods and Tools, Pages: 390-395, 2007.

作者在攻读硕士学位期间发表的论文与研究成果

发表的学术论文

1. Lele Liu, Liying Kang, Xiying Yuan, On the principal eigenvectors of uniform hypergraphs. Linear Algebra and its Applications, 511 (2016) 430-446. (SCI 收录)
2. Wei Zhang, Lele Liu, Liying Kang, Yanqin Bai, Some properties of the spectral radius for general hypergraphs. Linear Algebra and its Applications, 513 (2017) 103–119. (SCI 收录)

研究成果

1. Ahhy Lau, 上海大学研究生(硕博)学位论文 LaTeX 模板 ShuThesis.

致　谢

衷心感谢导师 xxx 教授对本人的精心指导.

附录 A　经典不等式

论文中用到的经典不等式.

(Hölder Inequality) 设 $a_i \geq 0, b_i \geq 0, i = 1, 2, \cdots, n$, 且 $p > 1, q > 1$ 满足 $1/p+1/q = 1$. 则有

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q \right)^{\frac{1}{q}},$$

等号成立当且仅当存在一个常数 c 满足 $a_i^p = c b_i^q$.

(PM Inequality) 设 $x_1, x_2, \ldots, x_n$ 是 n 个非负实数. 如果 $0 < p < q$, 那么

$$\left(\frac{x_1^p + x_2^p + \cdots + x_n^p}{n} \right)^{\frac{1}{p}} \leq \left(\frac{x_1^q + x_2^q + \cdots + x_n^q}{n} \right)^{\frac{1}{q}},$$

等号成立当且仅当 $x_1 = x_2 = \cdots = x_n$.

(AM-GM Inequality) 设 $x_1, x_2, \ldots, x_n$ 是 n 个非负实数. 则有

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n},$$

等号成立当且仅当 $x_1 = x_2 = \cdots = x_n$.