

武汉理工大学

(申请理学硕士学位论文)

微博社交网络的对称性研究

培养单位：理学院

学科专业：应用数学

研究生：康泽东

指导老师：余旌胡 教授

2014 年 12 月

微博社交网络的对称性研究

康泽东

武汉理工大学

分类号_____

密 级_____

UDC _____

学校代码 10497

武汉理工大学

学 位 论 文

题 目 微博社交网络的对称性研究

英 文

题 目 Study on symmetry of micro-blog social networks

研究生姓名 康泽东

指导教师 姓名 余旌胡 职称 教授 学位 博士

单位名称 理 学 院 邮编 430070

副指导教师 姓名 丁义明 职称 研究员 学位 博士

单位名称 中科院武汉物理与数学研究所 邮编 430071

申请学位级别 硕 士 学科专业名称 应用数学

论文提交日期 2014 年 10 月 论文答辩日期 2014 年 12 月

学位授予单位 武汉理工大学 学位授予日期

答辩委员会主席 评阅人

2014 年 12 月

独 创 性 声 明

本人声明，所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。据我所知，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经发表或撰写过的研究成果，也不包含为获得武汉理工大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中明确的说明并表示了谢意。

研究生签名：_____日期：_____

关于论文使用授权的说明

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后遵守此规定）

签名： 导师签名： 日期：

摘 要

对称性是微博社交网络中最直观也是最重要的性质之一，研究微博社交网络的对称性变化规律，可以揭示微博社交网络用户节点之间的联系紧密程度。然而，目前已有的社交网络对称性研究方法存在局限性。一方面，微博社交网络的对称程度没有具体准确的定义，容易忽略微博的特殊性；另一方面，现有的对称程度算法不适用于大数据类型，而社交网络研究属于大数据挖掘范畴。为了解决这两个方面的问题，本文基于微博的特性，给出微博社交网络对称程度的定义，并对已有算法进行优化，使得对大数据类社交网络对称性的研究更加可行。

本文主要分为三个部分，第一部分主要阐述微博社交网络对称性的相关背景，并概述了现有的Reciprocity算法，给出了改进的对称程度算法。第二部分使用改进算法研究微博社交网络的对称性，具体工作如下：根据收集的100万条新浪用户之间的关注关系和236个Twitter用户及其之间的关注关系来构建初始微博社交网络，选取其中具有明显对称性的连通子网络作为研究的主要对象；通过去除法研究影响其最大连通子网络对称性的主要因素，并分析了产生此现象的原因；用比较分析法，得到了Sina微博和Twitter的大V用户构成的社交子网络的对称性强弱关系；从功能定位方面给出了两种微博的对称程度实证分析；通过对初始网络的所有连通子网络的对称性研究，总结得出社交网络的对称性变化规律。第三部分讲述了社交网络的对称性变化规律的现实意义，总结了相关的结论，并对微博社交网络的对称性的进一步研究提出了延伸和展望。

关键词：微博社交网络；对称程度；连通子网络

Abstract

Symmetry is one of the most intuitive and important properties of micro-blog social network. By studying its symmetry's change regulation of micro-blog, the degree of the close link of micro-blog social network users among the nodes can be revealed. However, there are limitations of research methods for social network's symmetry. On the one hand, there is no specific accurate definition symmetric degree for micro-blog social network, ignoring the specificity of micro-blog easily. On the other hand, the existing symmetric algorithms do not adapt to the type of big data, while social network's research belongs to the areas of data-mining for big data. In order to solve these two problems, this article will give the definition of the micro-blog social network's symmetry degree, and the algorithm is also optimized to make the research of the symmetry of social network more feasible based on the features of micro-blog.

This article is divided into three parts, the first part mainly elaborated the relevant backgrounds of symmetry in micro-blog social networks, and summarized the existing Reciprocity Algorithm, and its improved symmetry degree algorithm. In the next part, the improved algorithm was applied to research micro-blog social network's symmetry, the specific works were as follows: Based on the collection of 1 million focusing relationships among the Sina micro-blog users and 236 twitter users as well as their focusing relationship, the initial micro-blog social network was established, choosing one sub-network with a clear symmetry as the main research object. By the elimination method, this paper studied the main factors which affected the symmetry of the maximum sub-network, and analyzed the causes of this phenomenon. A comparative analysis was used to find out the symmetry's relationships between the social sub-networks consisted of big V users in Sina micro-blog or Twitter. The empirical analysis of these two micro-blog's symmetry degree was given in terms of their functions. By researching the symmetry of all connected sub-networks of the initial network, social network's symmetry change regulation was summarized. The last part described the practical significance of the change rules of social network's symmetry, then summarized the relevant conclusions, its extension and outlook to the further study of the symmetry of micro-blog social network was put forward in the end.

Keywords: micro-blog social network; symmetry degree; connected sub-network

目 录

摘 要.....	I
Abstract.....	II
第一章 绪论.....	1
1.1 研究背景与研究意义.....	1
1.1.1 研究背景	1
1.1.2 研究意义	1
1.2 国内外研究现状.....	2
1.3 本文主要的研究工作.....	2
第二章 预备知识.....	4
2.1 数据来源与处理.....	4
2.2 相关理论介绍.....	5
2.2.1 微博社交网络图	5
2.2.2 图论相关概念简介	6
2.2.3 社交网络对称程度模型	6
2.2.4 重要节点的概念简介	9
第三章 最大连通子网络研究.....	15
3.1 微博社交网络图.....	15
3.2 影响最大连通子网络对称性的主要因素.....	18
3.2.1 大 V 用户和可忽略用户的定义	18
3.2.2 影响因素研究	29
第四章 大 V 用户构成的社交子网络研究.....	34
4.1 对称程度分析	34
4.2 两种微博社交网络的实证分析	35
第五章 初始网络的对称性变化规律研究.....	37
5.1 初始网络的对称性变化规律.....	37
5.2 现实意义.....	37
第六章 总结与展望.....	40
参考文献.....	41
致 谢.....	44
攻读硕士学位期间发表论文及参与项目情况.....	45

第一章 绪论

1.1 研究背景与研究意义

1.1.1 研究背景

微博作为一种新兴的社交媒体，其用户以及影响力越来越广泛，微博从一开始的社交娱乐工具到现在的重要营销手段，得到了前所未有的关注^[1]。把微博注册用户作为节点，相互之间的关注关系作为连线依据，这样便组成一个微博社交网络。它属于复杂网络的范畴，其拓扑性质的研究已经比较成熟，比如微博用户关系网络具有无标度特性和小世界效应^[1]。随着社会的发展，这些性质已经不能满足研究员的需要，一些其他专属于微博社交网络的性质渐渐浮出水面，社交网络的对称性就是其中一个重要的性质。随后，国内外的学者和研究员们对此性质展开分析研究，但是这些研究只是停留在对社交网络的对称性性质的研究，没有从量的角度深层次的挖掘，缺乏一个衡量社交网络对称性的指标。随着人们越来越意识到微博社交网络对称性研究的重要性，科研人员也开始关注如何寻找一个度量来衡量社交网络的对称性的问题。Reciprocity 是由 Social Media Mining-An Introduction,2014^[2]一书中提出的度量社交网络对称性的量，并给出了一种计算方式，但是此计算方法对于较大数据实现起来不易，为此本文提出一种更容易计算大社交网络对称程度的算法，简称社交网络对称程度算法。该算法将所要研究的社交网络转化为邻接矩阵的形式，通过做差计算出不对称用户，从而计算出社交网络的对称程度，大大增加了计算大型社交网络的对称程度的可行性。

1.1.2 研究意义

利用社交网络对称程度算法模型，我们可以计算出大数据网络的对称性变化规律，而这些对称性变化规律又可以反映出圈子联系的紧密程度。社交网络的基本准则：如果你关注了某些人，你希望他们同时反过来关注你^[3]。由于社交网络的不对称性质，该准则往往无法实现。相应的，究竟是什么在打破这种准则，其中谁又起主要作用，这些问题在以前都是迷。具体分为两种情况：对于一个社交网络而言，对称性随着圈子规模变化的规律，究竟是哪些因素在影响社交网络对称性，其中哪些因素起主要作用；对于多个社交网络而言，对称性随圈子规模变化的规律是否相同，其中影响社交网络对称性变化的因素是否相同，多个社交网络之间的比较等等。

1.2 国内外研究现状

社交网络作为大数据挖掘的范畴，已经得到了社会各界的广泛关注，其中微博^[4]是近些年来发展起来，渐渐取代博客的社交网络。其功能强大、操作简单让很多人慢慢习惯这种像日记般的记录自己生活的社交工具，人们的生活越来越离不开微博。注册用户在微博中的关注关系构成微博社交网络。

关于微博社交网络的拓扑性质的研究，其中尹书华^[5]以新浪、腾讯、搜狐三大微博注册用户的关注关系为研究对象，构建复杂网络，利用 PAJEK，MATLAB 等软件，计算分析该网络的度分布、聚集系数以及平均路径长度等统计特征，得出三大微博用户关系网具有无标度特性以及小世界效应，给出了微博用户关系网络的进一步研究方向。

关于社交网络用户对称性方面的研究，国内的尹志军等^[6]为了凸显网络中好友关系之间相互影响能力的非对称性，提出了 Partial strength 的概念补充了传统社交网络的 Tie strength 概念的不足；肖仰华^[7]详细阐述了图结构对称性的概念及其在数据管理问题的应用；吴文涛^[8]在此基础上详细讲述了图对称理论及其在社会网络分析若干问题中的应用。

国外的 Siegler^[9]以 Twitter 为例，提出了人们确定对方是否值得关注的“黄金比例”，指出 Twitter 是一个不对称的社交网络；Garrido^[10]提出 symmetry/asymmetry levels 的概念来刻画复杂网络的对称性。文献[2]提到社交网络的 reciprocity 是度量社交网络对称性的量，并给出一种计算社交网络的 reciprocity 的方法。

1.3 本文主要的研究工作

研究目标：微博作为当今社会炙手可热的信息交流分享平台，已经渐渐霸占我们的闲暇时间，从上班挤地铁、公交到外出应酬朋友聚会，人们无时无刻都在拍照上传分享自己的所见所闻，他们渴望得到其他人的关注。这也使得对微博社交网络的性质研究变得更加理所当然，大量的相关文章也在这个时候出现，也在同时本文提出了一种较简便计算大社交网络的对称程度的算法，用来观察其对称性随圈子变化的规律。这些规律可以用来分析现在主流社交网站的功能定位，在量的层面上解释一些在现实中看似成立但是又无法从科学的、理论的、系统的角度论证的问题。最后，我们了解到这些规律后，可以指导我们的日常生活习惯，避免一些社会问题的发生。

研究内容：本文首先对网络上搜集的数据进行转化、筛选处理，最终转换为 MATLAB 等软件可以识别的格式。其次构建微博用户关注关系初始社交网络，并选取其中具有较高对称性的连通子网络作为主要研究对象。再次在对微博用户关注关系初始社交网络的最大连通子网络分析的基础上探讨影响社交网络对称性的因素，并确定哪些因素起主要作用，哪些因素起非重要作用；对微博用户关注关系社交网络的所有连通子网络对称程度随社交圈规模变化的一些规律。然后在实证分析上，通过对 Sina 微博和 Twitter 用户关注关系社交网络的研究，分析得出其中的规律。最后分析比较 Sina 微博和 Twitter 的功能定位的不同及其产生原因。

本文主要创新点：在对原始数据的处理过程中，我们的选取方法是经过几次对比实验分析后得到的，它为后续的研究工作带来了可能。在构建微博用户关注关系初始社交网络的过程中，我们采用了 MATLAB 等编程软件，并使用 PAJEK 软件可视化微博社交网络，便于我们更加清楚的找到其中的连通子网络，易于理解究竟为何要选取初始社交网络的连通子网络。在对初始社交网络的最大连通子网络的研究中，我们创新性的定义了大 V 用户和可忽略用户，便于我们更加形象的描述影响最大连通子网络对称性的主要因素，关于找出大 V 用户和可忽略用户的方法，我们引入了重要节点的概念，使得这些定义更加具有说服力。

第二章 预备知识

2.1 数据来源与处理

(1) Sina 微博

Sina 微博数据来源于自然语言处理与信息检索共享平台^[11]（Natural Language Processing & Information Retrieval Sharing Platform, NLPIR）中的微博关注关系语料库。

```
<?xml version="1.0" standalone="yes"?>↓
<RECORDS>↓
<RECORD>↓
<person_id>10145</person_id>↓
<guanzhu_id>10029</guanzhu_id>↓
</RECORD>↓
<RECORD>↓
<person_id>10145</person_id>↓
<guanzhu_id>10318</guanzhu_id>↓
```

图 2.1 Sina 微博原始数据部分展示图

原始数据格式为 xml 格式，首先使用 xml2csv 软件转化为 csv 格式，csv 格式文件可以使用 MATLAB 软件编辑。转换好以后需要使用 notepad++ 软件打开，删除第一行得到 xinlang.csv 才可在 MATLAB 软件中正常使用。然后使用 MATLAB 软件选取既存在被关注列表中，又存在关注列表中的用户节点。100 万条关注关系数据当中有 1379 个这样的用户节点，再从 xinlang.csv 当中找出 1379 个用户节点之间的关注关系，发现存在 98 个孤立节点，这些节点与其他 1281 个节点之间没有关注关系。最后去除孤立点后得到初始 Sina 微博社交网络矩阵，其中用户节点数为 1281 个，节点关注关系数为 4417 条。

(2) Twitter

Twitter 数据来源于数据堂关于 Twitter 的 973 个 networks^[12]。首先将其中所有文件后缀名为 edges 的文件统一修改为 txt，通过实验结果对比筛选后选取其中的 12831.txt 矩阵作为研究对象。

```
398874773 652193
18498878 14749606
14305022 8479062
22253 12741
15540222 14809096
7415132 14172562
17129553 13839772
15911247 14086492
```

图 2.2 Twitter 原始数据部分展示图

使用社交网络语言，将 12831.txt 矩阵转化为初始 Twitter 社交网络矩阵。该矩阵中，用户节点数为 236 个，用户关注关系数为 2478 条。

2.2 相关理论介绍

2.2.1 微博社交网络图

社交网络中最基本的元素是点和线，其中点代表用户和行为者，线代表行为者之间的关系或者联系，点和线的联系构成图。微博社交网络中的节点是微博平台上的用户，连接微博社交网络中节点的连线表示节点之间的关系，在微博社交网络中这种关系是通过微博用户之间的“关注”“被关注”来表示的。微博社交网络是一种有向网络，即节点之间的关注关系并不是双向的，则一个节点关注其他节点并不代表其他节点一定关注该节点，因此关注者所发布的微博不能传递给被关注者^[13]。用 A、B、C、D、E 表示 5 个微博用户，他们之间的关注关系为 A 关注 B，B 关注 C 和 D，C 关注 D 和 E，D 关注 A、B 和 E，E 关注 D，则由这 5 个节点及它们之间的关系所构成的微博社交网络图，如图 2.3 所示。

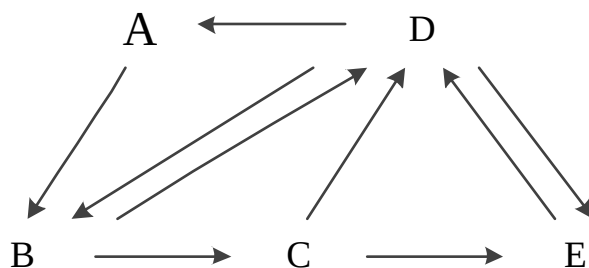


图 2.3 五个微博用户关注关系社交网络图

在这个微博社交网络中，用户节点数为 5 个，用户关注关系数为 9 条。

2.2.2 图论相关概念简介

(1) 有向图^[14]

有向图 D 是指有序三元组 $(V(D), A(D), \varphi_D)$ ，其中 $V(D)$ 是非空的顶点集； $A(D)$ 是不与 $V(D)$ 相交的有向边集；而 φ_D 是关联函数，使得 D 的每条有向边对应于 D 的一个有序顶点对（不必相异）。若 a 是 D 的一条有向边，而 u 和 v 是满足 $\varphi_D(a) = (u, v)$ 的顶点，则称 a 为从 u 连接到 v 的一条弧（或有向边），称 u 是 a 的始点， v 是 a 的终点，在不产生混淆的情况下，可简记有向边 a 为 uv 。

(2) 有向子图^[14]

对于有向图 $D = (V(D), A(D), \varphi_D)$ 。若 $V(D') \subseteq V(D)$ ， $A(D') \subseteq A(D)$ ，并且 $\varphi_{D'}$ 是 φ_D 在 $A(D')$ 上的限制，则称有向图 $D' = (V(D'), A(D'), \varphi_{D'})$ 是 D 的有向子图。

(3) 有向路^[14]

有向图 D 的有向路是指一个有限非空序列 $W = (v_0, a_1, v_1, \dots, a_k, v_k)$ ，其各项交替地是顶点和有向边，使得对于 $i = 1, 2, \dots, k$ ，有向边 a_i 有始点 v_i 和终点 v_{i-1} 。有向路 $(v_0, a_1, v_1, \dots, a_k, v_k)$ 常简单地用其初终点序列 (v_0, v_k) 表示。

(4) 单向连通有向图^[14]

若有向图 D 中存在有向路 (u, v) ，则顶点 v 称为在 D 中从顶点 u 出发可到达；若 D 中任意两个顶点 u, v ，顶点 u 可达 v 或 v 可达 u ，则称 D 为单向连通有向图。

2.2.3 社交网络对称程度模型

(1) Reciprocity^[2]

Reciprocity(交互性)是简化了的传递性，这是由于它考虑了长度为 2 的闭环，这是在有向图中才有的概念。正式地，如果节点 v 与节点 u 相连，则称 u 通过链接与 v 存在交互性。例如，在微博网站 Tumblr 中，“共同好友”就充当了这些节点。通俗的讲，Reciprocity(交互性)可以理解为，如果我成为你的朋友那么你也会是我的朋友。

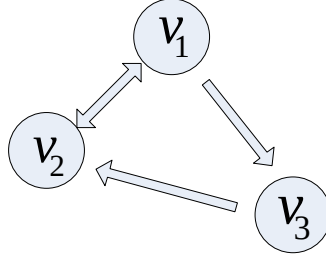

 图 2.4 v_1, v_2, v_3 之间的关注关系图

图 2.4 展示了图像中两个节点 (v_1 与 v_2) 之间的 Reciprocity (交互性)。

Reciprocity(交互性)可以计算出图中具有交互性的节点对的数目。在任意有向图中, 这些节点对的最大数目是 $|E|/2$ 。并且当有向图的所有边界都是交互的时候, 图具有的交互性节点对的数目达到最大, 为 $|E|/2$ 。因此, 节点对的数目值可以作为正则化的一个因素。利用邻接矩阵 A 可以计算出 Reciprocity (交互值)。

$$\begin{aligned}
 R &= \frac{\sum_{i,j,i < j} A_{i,j} A_{j,i}}{|E|/2}, \\
 &= \frac{2}{|E|} \sum_{i,j,i < j} A_{i,j} A_{j,i}, \\
 &= \frac{2}{|E|} \times \frac{1}{2} \text{Tr}(A^2), \\
 &= \frac{1}{|E|} \text{Tr}(A^2), \\
 &= \frac{1}{m} \text{Tr}(A^2),
 \end{aligned}$$

其中 $\text{Tr}(A) = A_{11} + A_{22} + \cdots + A_{nn} = \sum_{i=1}^n A_{ii}$, m 是网络中边界的数目。注意, 当所有的边界都交互时, m 是 $\sum_{i,j,i < j} A_{i,j} A_{j,i}$ 的最大值。

图 2.4 所示的图中邻接矩阵是

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

它的 Reciprocity（交互值）是

$$R = \frac{1}{m} \text{Tr}(A^2) = \frac{1}{4} \text{Tr} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \frac{2}{4} = \frac{1}{2}.$$

（2）对称程度

1) 社交网络对称程度定义^[15-18]

对于微博社交网络，有向图 $D = (V(D), A(D), \phi_D)$ ，其中 $V(D)$ 是用户节点的集合， $A(D)$ 是用户节点之间有向边的集合， ϕ_D 是用户节点关注关系函数。若 a 是 D 的一条有向边，而 u 和 v 是满足 $\phi_D(a) = (u, v)$ 的用户节点，则称 a 为从 u 连接到 v 的一条有向边，称 u 是 a 的始点， v 是 a 的终点。用户 u 是用户 v 的粉丝， u 称为关注者， v 称为被关注者。

① 任意两个用户之间可能存在的关系

a. 单向连线关系。它的含义是两个用户之间只有其中一个用户关注另一个用户，则两用户之间只存在一条单向的连线。例如用户 $U1$ 关注了用户 $U2$ ，但用户 $U2$ 没有关注用户 $U1$ ，那么 $U1$ 与 $U2$ 之间，只有一条 $U1$ 指向 $U2$ 的连线；同样， $U1$ 没有关注 $U2$ ，但是， $U2$ 关注了 $U1$ ，那么 $U1$ 与 $U2$ 之间，只有一条 $U2$ 指向 $U1$ 的连线；

b. 双向连线关系。它的含义是两个用户之间互相关注，则两用户之间存在一条双向的连线。例如 $U1$ 与 $U2$ 互相关注，那么 $U1$ 与 $U2$ 之间，有两条相互指向的连线。

② 对称用户与不对称用户

具有单向连线关系的用户称为不对称用户。

具有双向连线关系的用户称为对称用户。

③ 对称程度

社交网络对称程度是指所有对称用户之间的有向边数与图的有向边总数之比。

2) 社交网络对称程度算法

有向图 $D=(V,A)$ ，其中用户节点集合 $V=\{v_1,v_2,\dots,v_n\}$ ，用户节点之间有向边集合 $A=\{a_1,a_2,\dots,a_n\}$ ， $|V|$ 和 $|A|$ 分别是有向图 D 的用户节点总数和用户节点之间的有向边总数。

① 计算有向图的邻接矩阵

用 b_{ij} 表示从用户节点 v_i 到用户节点 v_j 的有向边的条数， b_{ij} 的可能取值为 0,1，其中 v_i 和 v_j 为 D 的用户节点， $B(D)=(b_{ij})_{n \times n}$ 为有向图 D 的邻接矩阵；

② 计算有向图的邻接矩阵中所有不对称用户之间的有向边数

对于有向图 D 的邻接矩阵 B ，做如下运算： $B-(B+B')/2$ ，记作 N ，即 $N=B-(B+B')/2$ 。此时，矩阵 N 的元素由 0.5,-0.5 和 0 组成。为了计算上的简便，将矩阵 N 所有值为 0.5 的元素替换为 1，将所有值为-0.5 的元素替换为 0，得到一个新的元素由 0 和 1 组成的矩阵 M 。矩阵 M 中值为 1 的元素总数便是所有不对称用户之间的有向边数，记作 $\text{sum}(M(:))$ ；

③ 计算最大连通子网络的对称程度

用 $|A| - \text{sum}(M(:))$ 做差便得到最大连通子网络的所有对称用户之间的有向边数。于是根据对称程度的定义，得到对称程度 (P) 的计算公式：

$$P = (|A| - \text{sum}(M(:))) / |A|$$

注明：其它社交网络亦可用此算法计算其对称程度，只需要将相应的数值替换成要计算的社交网络的数值。

2.2.4 重要节点的概念简介

(1) 有限马尔可夫链相关性质^[19-21]

1) 本文研究的马尔可夫链特征

X_n 的所有可能取值为系统的状态，一个离散时间的随机过程 $X_n(n=0,1,2,\dots)$ ，其中 X_n 在有限集合 $S=\{1,\dots,N\}$ 或者 $S=\{1,\dots,N-1\}$ 上取值。

为了描述这个随机过程的概率性质，我们首先给出它的概率

$$P\{X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\}, \quad (2-1)$$

其中 n 为任意整数, $(i_0, \dots, i_n)$ 为任意有限状态序列。

其次, 我们给出初始概率分布状况

$$\phi(i) = P\{X_0 = i\}, i = 1, \dots, N \quad (2-2)$$

以及转移概率

$$q_n(i_n | i_0, \dots, i_{n-1}) = P\{X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}\}, \quad (2-3)$$

(2-2) 式、(2-3) 式是由于

$$P\{X_0 = i_0, \dots, X_n = i_n\} = \phi(i_0)q_1(i_1 | i_0)q_2(i_2 | i_0, i_1) \cdots q_n(i_n | i_0, \dots, i_{n-1}). \quad (2-4)$$

在文本的研究中, 我们只涉及到一类特殊的随机过程--满足马尔可夫性的随机过程。马尔可夫性, 即为 (2-5) 式,

$$P\{X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}\} = P\{X_n = i_n | X_{n-1} = i_{n-1}\}. \quad (2-5)$$

再次, 我们假设该随机过程满足时齐性, 即转移概率与时间无关。时齐性, 即为 (2-6) 式

$$P\{X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}\} = p(i_{n-1}, i_n) \quad (2-6)$$

的过程, 其中 p 是 $S \times S \rightarrow [0, 1]$ 上的函数。

2) 本文研究的马尔可夫链性质

我们可以由初始概率分布和转移概率得到本文研究的马尔可夫链的概率性质。首先, 本文研究的马尔可夫链的概率可化简为

$$P\{X_0 = i_0, \dots, X_n = i_n\} = \phi(i_0)p(i_0, i_1)p(i_1, i_2) \cdots p(i_{n-1}, i_n). \quad (2-7)$$

其次, 本文研究的马尔可夫链的转移矩阵 P 是一个 $N \times N$ 矩阵, 其 (i, j) 位置上的元素 P_{ij} 为 $p(i, j)$, 并且, P 是一个随机矩阵, 它满足

$$0 \leq P_{ij} \leq 1, 1 \leq i, j \leq N, \quad (2-8)$$

$$\sum_{j=1}^N P_{ij} = 1, 1 \leq i \leq N. \quad (2-9)$$

任意满足 (2-8) 式, (2-9) 式的矩阵均可作为本文研究的马尔可夫链的转移矩阵。

3) 在给定状态下, 计算马尔可夫链概率的步骤

当已知转移矩阵 P 以及初始概率分布 ϕ 的情况下, 计算马尔可夫链在时刻 n 处于状态 j 的概率的步骤如下:

首先, 定义 n 步概率 $p_n(i, j)$ 为

$$p_n(i, j) = P\{X_n = j | X_0 = i\} = P\{X_{n+k} = j | X_k = i\} \quad (2-10)$$

(2-10) 式的第二个等式的成立是根据 (2-6) 式得来, 从而

$$P\{X_n = j\} = \sum_{i \in S} \phi_i P\{X_n = j | X_0 = i\}. \quad (2-11)$$

其次, 我们来证明 n 步转移概率 $p_n(i, j)$ 实际就是矩阵 P^n 在 (i, j) 位置上的元素。第一步, 对于 $n=1$, 这一结论明显成立; 第二步, 假设对任意给定的 n , 此结论成立, 即

$$\begin{aligned} P\{X_{n+1} = j | X_0 = i\} &= \sum_{k \in S} P\{X_n = k | X_0 = i\} P\{X_{n+1} = j | X_n = k\} \\ &= \sum_{k \in S} p_n(i, k) p(k, j). \end{aligned} \quad (2-12)$$

成立。第三步, 由第二步知, $p_n(i, j)$ 是 P^n 在 (i, j) 位置上的元素, 所以最后求得和就是 $P^n P = P^{n+1}$ 在 (i, j) 位置上的元素。

再次, 一个马尔可夫链的初始概率分布可以通过向量

$$\bar{\phi}_0 = (\phi_0(1), \dots, \phi_0(N)) \quad (2-13)$$

给出。我们用 $\bar{v}$ 表示向量 $(v(1), \dots, v(N))$, $\bar{v}$ 既可以是行向量也可以是列向量。

最后, 假定 $\bar{\phi}_0$ 已知, 那么在时刻 n 的分布 $\phi_n(i) = P\{X_n = i\}$ 可以表示为

$$\bar{\phi}_n = \bar{\phi}_0 P^n. \quad (2-14)$$