硕士学位论文

社交网络节点及其关系的研究

Research on nodes and relationship in social network

作者姓名：吴 慧

学科、专业：计算机应用技术

学号：21209203

指导教师：张绍武

完成日期：2015-05-01

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年_____月_____日

摘　要

在线社交网络的快速发展，冲击和改变着人们的生活和思维方式。于无形之中，人们已成为各种社交网络中的一员，比如，在各种诸如 QQ，微博等通讯社交网络中，人们发表着自己的生活琐事，抒写着自己的心情故事。在社交网络中人们不仅仅表达自己的生活，同时也洞察周围的人和事。对社交网络的研究也具有重要的实际意义和理论意义，比如有针对性的研究社交网络中某一个用户，可以有的放矢的推荐产品、推荐朋友、或者引导某种舆论观点的传播等。如何将具体的社交网络抽象成网络拓扑图，研究网络图的节点和节点关系正是本文所讨论的主题。

不同的社交网络结构图中，节点和边代表不同的实际意义，比如在 Facebook 社交网络中，节点表示用户，边表示用户之间的朋友、转发等关系。本文也考虑了不同社交网络的特点，主要的研究工作从两个方面展开。

首先，社交网络节点的研究。社交网络中，每一个节点的重要程度不一样，其影响力也不尽相同，对节点影响力的度量可以帮助用户或商家找到影响力大的用户。本文提出了一种在微博社交网络中评价用户的影响力的算法，和传统的用户影响力的评价方法相比，综合考虑用户的活跃度和用户所发微博质量两个方面的因素。通过在公开语料集和真实数据中的实验，表明该方法是可行的，并比传统的用户影响力的评价方法更能客观真实的反映用户的实际影响力。

其次，社交网络节点关系的研究。如何预测社交网络中节点的链接情况，本文从节点属性和网络拓扑结构两个方面对节点的未知链接进行研究。在对以往传统方法基于节点邻近度和基于网络拓扑结构的方法总结后，提出了基于 AdaBoost 提升算法对未知链接进行预测。实验在公开的数据集科学论文合著网络上进行，和传统的基于节点属性和基于路径的方法等单一的弱分类器都进行了比较，证明基于 AdaBoost 方法的强分类器能更准确的提高未知链接的准确率，提高分类性能。

最后，对社交网络中节点及其关系的研究进行了归纳和总结，并提出了对未来工作的展望。对社交网络的研究，可以从不同的侧重点进行，每一种算法都有其不足和改进之处，对社交网络的研究在实际的应用中具有指导作用。

关键词：社交网络；节点影响力；链接预测；网络

Research on nodes and relationship in social network

Abstract

With the rapid development of online social networks, the way of life and thinking has been impacted and changed. In the invisible, people have become a member of various social networks. People publish their things of life and express their own story in some communicative social networks such as QQ and micro-blog. Not only can they express their life, but also can observe people and things around. The research on social network has important theoretical and practical significance. For example, it has a definite object in view of recommend products or friends or to lead some opinion, if a targeted study on a certain user on social network has been conducted. The theme in this article is the nodes and the relationship between nodes of a social network.

The nodes and the edges stand for different practical meanings in different social network structure. For example, the nodes represent users and the edges the relationship of friends or forwards in the social network of Facebook. This paper takes different characteristics of different social networks into consideration, and the study spreads as the two aspects.

First, the research on nodes in social network has been conducted. In social networks, the importance of every nodes is different, so as the influence. The method of how to measure the significance of nodes helps the users or the merchant find nodes having the greatest influence. The paper puts forward an algorithm to evaluate the influence of users in social networks. Compared with traditional methods, the paper thinks over the active of the users and the quality of the micro-blogs. The experiments on both public data sets and real data show that the method is feasible, and it can reflect the real influence of users better compared with the traditional methods.

Second, the research on the relationship of nodes in social networks has been conducted. The study develops from the two aspects: the attributes of nodes and the topological structure of the network. After a conclusion to traditional methods on node proximity and topological structure, a method based on AdaBoost to predict unknown links has been raised. Experiments on public co-authorship data sets prove that the strong classifier based on AdaBoost has more accurate to improve the accuracy of the unknown links, compared with the traditional weak classifiers based on traditional node attribute and path.

Last, induction and summary on social networks has been conducted, and future work is put forward. The study on social network can be developed from different points, and every algorithm has its own weakness. The research on social networks has a guiding role in actual application.

Key Words: social networks; node influence; link prediction; the topological structure

目　　录

1 绪论

1.1 研究背景

在网络越来越发达的今天，每个人都无形的包围在各种社交网络中，对社交网络的研究也引起了学者的兴致。不同的社交网络有不同的侧重点，人们也在不同的社交网络中发表着自己的声音，在 QQ 等社交网络中更多的是私人通信；微博中，博客们即性的发表自己的生活点滴；豆瓣等社交网络方便了人们日常的购物需要等等。这些看似迥然不同的社交网络，能否找到从特殊到一般的共性，也是研究员们反复猜测和证明的命题。发现社交网络的结构特征，社交网络中用户和用户关系的内在逻辑关系，对于实际网络具有指导和借鉴意义。

在对社交网络结构的分析中，有很多针对社交网络普适性的特征被广泛的应用，很多适应于复杂网络、无标度网络的特征也都被应用到传统的社交网络中。将社交网络的特征进行抽象，从数值的角度去分析社交网络。社交网络图分为有向和无向的，反映了网络节点对之间的关系是否为对等的；社交网络中度的分布反映了网络中单个节点与整个社交网络的联系；聚类系数反映了网络松散耦合的程度；平均路径长度也对小世界网络中“六度分割”理论[1]的证明。然而，在对不同的社交网络处理时，都是从社交网络的共性上研究，没有根据不同的社交网络，求同存异的个性化研究。本文主要在特定的社交网络中，从社交网络的组成部分，即节点和边，去分析社交网络的特征。文章主要从社交网络中节点及其关系两个方面进行细粒度的研究。

在对社交网络的研究中，是将社交网络抽象成由节点和边构成的具体的网络图形，分析网络图形中节点和边的情况，从图论的角度去分析，但是最终得到的结果是对一般社交网络的大致分析。而不同的社交网络具有不能的特征，比如在微博社交网络平台中，就具有评论、转发等特定的用户行为，所形成的网络边也可以由转发、粉丝等表示。

在社交网络图中，节点一般表示用户，在具体的社交网络中，每一个节点所处的地位并不一样，比如在微博社交网络中，每一个用户的影响力都不一样；在社区社交网络中，每一个用户的支持力也不相同；在媒体社交网络中，不同的媒体具有不同的影响力等等。如何去评判节点在其所处的社交网络中的具体地位，可以更精准的找到目标节点。

在对社交网络中节点影响力的分析中，无论是从网络图的结构特征，还是从网络拓扑结构来分析，都可以在一定程度上反映用户的影响力情况。但是，根据网络图的不同特征，“影响力”可以有不同的表达方式，不同的应用也侧重于不同的方面，从不同的角度去衡量，节点的“影响力”也在发生变化。在特定的社交网络中，怎样去诠释“影响力”，并没有一个定性的说明。而且，在特定的社交网络中，衡量影响力指标也不一样，最后计算得到的影响力大小，也需要有一个合理的度量方法。

对社交网络中节点影响力的评估，也有各种衡量方法。David[2]加入用户行为的因素，认为社交网络中用户的影响力由用户本身以及用户和其它用户的交互形成，即用户的自身属性和社交属性共同决定用户的影响力。这也为研究社交网络的节点影响力提供了一个思路，节点的影响力需要从自身和外界等多个方面综合考虑。比如微博社交网络中，一方面，微博用户具有特定的评论、转发等自身行为特征，反映了用户的参与热情，另一方面，微博中用户可以被粉丝跟随，其所发微博信息的被评论、被转发次数等，反映用户和其它用户的互动情况，互动越频繁，其信息也被传播的更快，具有口碑效应。在具体计算微博社交网络中用户影响力时，需要从不同的角度去考量，更加全面的得知该用户的影响力情况。最终得到的结果也需要在给定的评价体系进行评判。

在对用户节点的研究中，不仅仅需要衡量用户在某一个社交网络中的影响力，用户和用户之间的沟通和交流也具有重要的价值。社交网络的发展也改变了人们传统的交流方式，在社交网络中，人们不仅仅可以联络老朋友，还可以结交新朋友。现有的老朋友可以看作是已有节点链接，未来的新朋友可以看作是未知链接，预测未知链接，可以实现社交网络中的推荐、预测等。比如在生物蛋白社交网络中，发现未知的蛋白质组合，可以帮助创造巨大的价值。在对未来的链接预测中，发现现在没有链接关系，但是在未来可能有链接关系的节点对，可以很好的预知未来。

在对社交网络中节点关系的研究中，需要考虑两个或者多个节点的连接关系，主要是对未来的链接预测的分析。传统利用社交网络中节点相似性[3]进行度量的方法，提供了链接预测的基本思想，但是对链接预测的考虑比较单薄，节点相似性主要是在节点局部的网络中得到的链接预测算法，并没有考虑全局的网络结构，文章提取了某社交网络中节点的自身属性，并分析其在社交网络中的结构特征，预测在未来的时间内，可能出现的节点对。

在社交网络方兴未艾的发展中，不仅仅需要将社交网络抽象出来，从社交网络的共有特性去研究，也需要从个性化的角度分析每一个社交网络的不同特征。文章主要是针对社交网络的组成部分进行了细致的分析，从更细致的角度去分析社交网络的特点。在对社交网络中节点的研究中，找到特定社交网络中每一个节点的不同作用，找到该社交网络中影响力较大的用户，具有重要意义。另外，在对社交网络中节点间相互关系的研究中，预测未来时间段内与用户最有可能链接的节点。

对社交网络中节点及其关系的理论研究，可以将特定的社交网络抽象成网络拓扑图，赋予网络图中每一个节点和边不同的实际意义，将对实际的社交网络的研究转换成对网络拓扑图的研究，估量不同的节点影响力情况，节点之间的链接关系与否等，对实际的社交网络具有现实指导意义。

1.2 研究现状

1.2.1 个体影响力分析

对社交网络中用户影响力的研究，大体上有以下几个方面。

从考虑用户属性或用户所在社交网络中形成的网络结构角度，可以有以下两个方面。

（1）在用户的静态属性研究方面，Meeyoung Cha 等[4]在对 Twittr 社交网络中用户影响力评价的研究中，从用户的粉丝数，用户的转发数等排名来研究用户的影响力，其中粉丝数较多的一般为知名公众人物，反应了现在的流行趋势，转发数较多的一般是媒体，其微博的内容质量较高。但是该方法所得到的排名只是根据粉丝数或转发数等一项指标得到，没有综合考虑包括用户行为的一系列指标，使得排名有一定的限制，不能普遍的反映用户的实际影响力。考虑到微博是一种互动的社交网络，石磊等[5]提出了用户活跃度模型，通过考虑用户粉丝，用户发微博的频率等计算用户的活跃指数，从而得到用户的活跃度排名，用户活跃度虽然可以作为衡量微博用户影响力的一项指标，但并不能说明用户在微博中参与积极性越高越活跃，该用户就越具有影响力。Danah Boyd 等[6]将 Twitter 用户的转发、回复等行为表示成用户的行为权重，在权重的基础上计算用户的影响力。该文章对用户的行为考虑的较周全，但是又忽略了用户的粉丝数等因素。张华平，孙梦姝等[7]通过分析用户所发的微博数，粉丝数，用户关注数的数值特征，得到用户的影响力模型。但是该论文得到的是用户群体的影响力，比如加 V 用户的影响力比普通用户的影响力高，没有得到个体用户的影响力。

（2）从社交网络中形成的拓扑结构的角度来研究，Yuto Yamaguchi 等[8]通过分析用户之间的关注关系，评估每一个用户在社交网络中的影响力。但是该方法仅仅考虑用户被关注边的多少，在网络图中所考虑的因素太少。Weng 等[9]基于 PageRank 算法根据用户之间的粉丝联系所形成的网络关系，得到用户的影响力排名。该方法用粉丝数量和质量来衡量用户的影响力较为直观，实现也较为简单，在 Twitter 社交网络中取得较好的效果，但实际的微博社交网络中存在僵死粉等现象，粉丝作为微博用户影响力的评价指标并不是很全面。王琛，陈庶樵[10]在改进了传统的 PageRank 算法的用户影响力评价方法，引入了微博传播能力这一概念，加入用户的行为可以更真实的反映用户的影响力。但是该方法提出的微博传播能力只包括了用户评论等很少量的信息，有一定的缺陷。徐健[11]通过研究复杂网络中的节点影响力，提出了新的中心性度量模型，不仅仅度量了节点的内部属性，也度量的节点的外部属性，得到的节点影响力模型可以衡量各种复杂网络包括社交网络的节点影响力。但该算法需要在不同的网络结构中，选取不同的节点传播源进行仿真，从采集数据到多组仿真实验，所耗用的时间空间较高。

针对以上两个方面的主要研究，不同的影响力因素有其不同的侧重点，也有将两者结合起来考虑，将用户的属性和用户所在社交网络中的拓扑结构同时进行研究，所得到的影响力较为普遍的反映用户的影响力[12,13]。

从社交网络中有向无向图，加权不加权的角度，有以下研究。

无向社交网络主要是指网络图中节点的连接没有方向性，比如节点 A 和 B，A 连接到 B 和 B 连接到 A 没有区别，在实际的网络中有论文合著网络，将作者看作是一个节点，作者之间的合著表示一条连边，所形成的网络图是无向的。而在微博等社交网络中用户之间的连接有方向性。在度量社交网络中节点影响力时需要从有向图和无向图等根据社交网络的不同特征考虑。另一方面，社交网络中的连边可以被赋予表示不同实际意义的权重，权重的大小也左右着节点的影响力情况。可以根据从一般到特殊的方法，加入社交网络的不同特点，度量不同的社交网络中节点的影响力。

谢世娜[14]等在发现影响力最大的节点算法中，提出了新的用户影响力度量思路，刻画每一个用户的影响力情况，并认为如果一个用户周围都是影响力较大的用户，那么该用户也具有较大的影响力，并分别给出了在无向和有向的社交网络中度量节点影响力的方法。比一般的贪心算法具有优势，但是该方法更多的是从网络自身情况去度量节点影响力大小，并没有结合社交网络中节点固有的属性进行分析。郭静[15]等将用户在社交网络中的影响力进行累计，提出了影响力传播的权重计算方法，以最大似然估计的方法，得知用户传播中的影响力最大化结果。但该方法需要考虑用户影响力的累计，需要取得用户在社交网络中历史的日志做为分析的样本数据，一个用户的影响力取决于与其相关联的用户的影响力情况，需要对每一个用户都分析样本数据，建立影响力模型，每个模型又有不同的特征，故所得到的用户影响力对模型、对样本的要求太高。

以上对社交网络中用户影响力的研究提供了方向，在微博社交网络中，用户影响力的研究也有其特性。微博中，用户的影响力由众多因素决定，如何将制约用户影响力的因素数据化，也是需要解决的首要问题。

卢体广[16]等抓取微博数据，建立数学模型，计算微博用户的影响力。其中，数学模型中主要考虑用户粉丝、发帖、评论等因素，文中给出了得到微博用户影响力因素量化的方法。但文章仅从技术的角度探讨了怎样得到微博数据，给出了一个通用的抓取数据的模型，数值化的分析微博用户的影响力，而没有考虑用户所在网络圈的影响力情况。黎明，文海英[17]等主要从微博用户的行为出发，评估用户影响力情况。在微博社交网络中，每一个用户具有不同的行为习惯，比如评论、转发、发帖等，结合用户的行为数据，在仿真实验中，以最小二乘方学习方法，说明用户行为对其影响力的作用。Romero[18]等度量 Twitter 社交网络中用户影响力，颠覆了以往对 Twitter 等社交网络中用户影响力的观点，即粉丝越多，影响力也越大，而是从两个不同的角度评估了 Twitter 用户的影响力：用户所发布的信息可以得到更多人的响应和用户的消极性。结合这两个方面的，提出了一个用户被动影响力的模型，可以比较准确的衡量 Twitter 社交网络中的用户影响力情况。微博可以看作是国内的 Twitter，但是也有所区别，算法可以为微博社交网络中用户影响力提供思路，但并不能移植到微博中，也需要加入微博自身的特定。Ding[19]

等针对微博中用户之间的互动行为，认为用户之间可能存在多个关系，进而形成多关系网络，在多关系网络中利用随机游走的度量模型计算每一个用户的影响力。

最新研究中，考虑到每一个社交网络中不同领域的人具有不同的影响力，针对某一个具体的话题，社交网络中的每一个用户对此的反映并不一样，进而得知针对某一话题的有影响力的用户群体。比如在对舆论控制中，可以发现具有舆论导向的用户群体，或者发现散布流言的用户群体等。刘继[20]等在微博社交网络中，挖掘不同主题下所得到的用户影响力，但是该算法主要是基于网络特征得到的影响力度量模型。没有考虑用户自身的属性，比如粉丝数等。吴渝[21]等为识别网络中的意见领袖，结合用户自身的属性特征和用户之间的连接关系，提出了识别意见领袖的算法，作者在采集不同时间段内，意见领袖的变化情况，更符合现实的需要。“意见领袖”是指社交网络中对舆论传播具有重要作用的节点，而有影响力的用户可能不是意见领袖，但是也可能对信息的传播具有重要作用。

通过以上各种算法的研究，本文综合考虑反映用户影响力的指标：用户活跃度和用户所发微博的质量。用户所发微博积极性越高，所发的微博质量高，微博将被广泛的转发和评价，其信息也相应的在网络中传播的更快，其影响力也越大。用户的积极活跃性考虑了用户在某一时间段内所发原创微博的频率，转发的频率，用户所发微博的质量考虑了用户所发微博在某一时间段内被转发次数和评价次数，得到用户的权重，该权重反映了用户的影响力，和传统的用户影响力评价指标比较，本文所提出的加入用户权重的影响力排名更具合理性，并能更客观真实的反映微博用户的实际影响力。

1.2.2 节点关系预测

在有关节点关系的预测中，文章主要讨论的是社交网络中节点链接与否的关系，即对链接预测的研究。

一般的研究内容主要从两个方面展开：对节点属性的研究和对网络拓扑结构的研究。

（1）对节点属性的研究。David 和 Joe 研究了网络中各种链路预测的方法[22]，文中提出的很多算法迄今被广泛的应用，改进和对比分析。在对链接预测的研究中，作者主要提出了三大类方法，第一类主要是根据节点的邻居节点得到的链接预测方法。在这种方法中，节点对的共同邻居数和节点对未来的链接概率呈现正相关。共同邻居数，Jaccard 系数，Adamic/Adar 算法等作为该类方法的链接预测算法。第二类主要是根据节点所在的网络路径得到的链接预测方法。这类算法的思想是，节点之间的网络路径和节点对链接的概率呈反相关。通过 Katz, PageRank, SimRank 等方法对网络的“最短路径”赋予了不同的预测方法。最后作者从数学的角度，将节点及其节点对看作矩阵的形式，从节点对所形成的矩阵的秩的角度研究节点可能链接的概论。最初的这篇论文所考虑的因素比较单一，但也给后来的学者提供了很多思路和借鉴的意义。在该论文的基础

上，很多相应的算法也进行了改进，Liyan Dong 和 Yongli Li 等[23]将网络的节点属性和网络拓扑结构结合形成新的算法，比起传统的单一的链接预测方法具有较好的预测效果。但是作者主要是将 Katz 和节点度链接起来，而没有考虑其他的链接预测方法，怎样的链接预测算法组合最优也是后来的学者不断研究实验的课题。

在以上基本的方法中，链接预测都是在邻居节点的基础上进行改进和变形实现，计算的复杂度较小，得到的链接预测信息也不够全面，相应的计算精确率也有待提高。吕琳媛，周涛等[24-27]对链接预测进行了各个方面的研究，在基于节点，基于路径等链接预测算法上都进行了改进。在基于邻居节点的方法上，作者提出了资源分配的链接预测算法 RA[28]，该算法虽然和 Adamic-Adar 算法有相似的形式，但是却取得了较好的准确率。考虑局部路径对链接预测的影响，作者将路径长度为 2 和 3 的节点对进行了预测，得到局部路径算法 LP,并取得了很好的结果。在对网络拓扑结构进行研究预测节点链接方面，各种随机游走算法也被提出。局部随机游走，叠加随机游走[29]均在局部路径的基础上提高链接预测的准确率。

（2）在网络拓扑结构方面的研究。对网络拓扑结构的研究也包括改变网络的结构，包括将整个社交网络图改变成一个二分部的图形式[30]；给节点或边赋予不同的权重[31]，形成一个加权的网络拓扑结构等等；通过对边赋予不等的权重，表达预测的不同需求[32]等等。张珊靓[33]等结合时间特征，并通过共同的话题将网络进行加权，在微博社交网络中，基于随机游走的算法，预测微博用户的未来好友推荐。文章主要分析用户在当前或最近的时间段内，和用户可能因某一话题而发生的好友链接关系，所设定的条件太多，并不具有社交网络中链接预测的一般普适性。郭景峰[34]等为了从网络的动态结构研究社交网络中节点的链接情况，将动态时间变化的序列和网络结构特征融合，提出了动态网络中的链接预测算法，该算法比起单独对静态属性的利用更优越。但是在将时间序列和网络结构进行融合时，只是进行了简单的乘法，比较单薄，没有复杂、可靠的统一模型。Jiawei Zhang[35]等通过将社交网络构造成一个异构的网络图，向用户推荐产品或好友。对传统的网络结构图进行了重构造，使得推荐的项目不局限于同质的人或物，但是在大规模网络中，构造异构型的网络本身就是一个较复杂的过程，使得对该算法的应用并不广泛。Zhaochen Guo[36]等根据随机游走提出鲁棒性的链接预测算法，文章主要从情感语义相似度的方面度量不同节点实体的相关性，节点之间的相关性影响着节点的随机游走概率，而游走概率较大的节点对，链接的几率也越大。分析节点的情感相关性大多从文本的角度挖掘用户的情感倾向，而单个用户的文本标签较少，需要爬取大量有关用户属性的语料，也必将增加实验的负担。潘果[37]等将用户的地理位置信息和用户所在的网络拓扑结构融合，给用户按照地理位置分类，推荐不同的好友。这也具有实际的意义。比如，通常用户更倾向于和在同一个地方的其它用户成为朋友，同理，在同一个地方的用

户也更有机会连接在一起。然而在社交网络中，本身相互链接的节点可能已经包含了地理位置信息，在针对特定的社交网络时，可能会有重复的考虑因素。李旺龙[38]等根据用户的不同质量，预测社交网络中用户之间是否关注的链接关系。在微博社交网络中，根据所形成的异构网络中不同节点的关系，将用户划分成不同的质量，预测不同用户群体的用户预测情况。该算法可以避免传统的链接预测中的局部性缺陷，但是用户质量问题带来一定的主观性。

在对链接预测的方法中，各种统计学习的方法也被广泛的应用，基于监督学习和半监督学习的方法[39-41]，也有基于马尔科夫链的学习方法[42]。通过这些方法在训练数据集中找到节点的属性特征或者网络特点，进而预测训练集中节点的链接状态。张玉芳[43]等利用马尔科夫逻辑网，构建关系模型，并预测网络链接与否。该方法可以对数据集中的噪音数据起到抗干扰的作用。然而该方法的应用对信息检索等具有重要作用，在诸如合著网络等数据集所代表的领域中还未得到广泛的应用。

链接预测也在不同的社交网络中具有不同的实际意义。比如在网络推荐系统中的链接预测方法，需要根据用户所在社交网络中的行为习惯，预测其在未来的时间段内可能出现的行为方式。在各种电子商务领域，可以从协同过滤[44]的角度考虑用户之间的项目相关性或者用户兴趣爱好相关性进而进行推荐等。

在以往的研究中，都是对一个算法进行了改进，每一个链接预测都有其理论基础和研究的背景，也为后续的研究工作和应用领域奠定了思路和方向。

在最新对节点关系的研究中，具有新的技术、新的思路、大数据量或加入新的元素等都有新的研究进展。朱索格[45]等根据目前网络中涌现的海量数据，提出了针对大量数据的数据抽样方法，使得对节点关系的研究可以从小范围的数据集扩展到大量的数据中进行，但是该方法主要是从节点之间的相似性出发，没有考虑节点所在的网络结构，同时大数据中的节点关系多样，根据节点网络关系建模难度太大。仇丽青[46]等通过改进邻居节点相似度，并加入时间的因素，分析随着时间的变化迁移，节点链接情况，并说明了节点的链接而导致的网络演化情况。作者给出了很好的融合时间的观念的方法，可以有效的去除噪音数据，但是对网络结构演化的过程，没有直观的给出数据的说明。卢文羊[47]等考虑在不同的社交网络中，节点本身具有不同的属性特征，比如具有节点的文本信息等，提出了基于 LDA 主题模型的链接预测方法，将节点文本内容相似性和邻居节点相似性结合，衡量在不同的文本内容的条件下，节点的相似度，最后找到最相似的第 k 的节点进行预测或推荐。也有一些研究针对节点的链接紧密程度。在社交网络中，将与节点联系紧密的称为“强链接”，而与节点联系稀疏的即为“弱链接”，由各种应用领域，衍生了各种对强链接和弱链接的新的研究领域。Giulio Cainelli[48]等从应用的角度说明了强链接的作用，但是算法的应用领域是论文合著，不具普适性。Philipp Mayr.Andrea Scharnhorst[46]在科学计量学和信息检索领域对弱链接进行了研究，对弱

链接的理解主要从不同领域的知识迁移展开。同时，也有对链接预测的新思路，从以往的预测链接到最新预测节点正面链接和负面链接。正面链接主要是指与节点相似或正相关的节点链接，负面链接主要是指与节点负相关的节点链接。Jiliang Tang[47]等主要从负面链接的角度进行了研究，负面链接对社交网络中的链接预测有积极的促进作用。Jerome Kunegis[48]等将链接分为节点相信的链接和节点不相信的链接，对两类链接进行了显著的标识，可以从正面和侧面等不同的角度去了解节点用户。在具体的应用中，可以更多的显示与用户相近或相信的用户信息，屏蔽与其兴趣相背的用户信息等。

从实际的应用来说，精准的链接预测涉及到方方面面，比如心理学，社会学，数据挖掘与分析等等，在著名的 Target 案例[49]中可以看到链接推荐的战略方法分析。而在数据化的研究中，怎样评价一个好的链接预测通常会有以下几个方面的考量。

- 同质性。一般认为，两个拥有更多相同属性的节点对更容易在未来发生链接，即“志趣相投”的朋友更可以成为朋友。
- 稀有性。两个节点对拥有稀缺的属性和节点对联系的概率呈正相关。比如，如果两个朋友都喜欢游泳，而另外两个朋友更喜欢极限运动，相比较，后者发生链接的概率更大。
- 社交网络的影响。如果很多人都有相关的属性，那么节点容易受社交网络的影响也拥有该属性。
- 共同的朋友。如果两个节点拥有更多的朋友，那么他们更可能在未来相遇成为朋友。
- 社交距离较近。潜在的链接一般在距离上较近。距离较近也让他们有更大的几率相遇。
- 偏好链接。人们更倾向于链接有影响力的人，比如刚进入一个新领域的社交网络中，用户更偏好与链接粉丝数较多的用户节点。

虽然每一个链接预测方法都针对不同的应用环境，不同的链接预测算法其侧重点也不相同，但是一个好的链接预测一般满足以上几点标准，如果把节点对之间的链接关系看作分类关系，分为有链接或没有链接，那么每一个链接算法都可以看作是一个分类器，预测节点对的链接关系。

为了充分的考虑基于用户节点和网络路径等影响节点链接的因素，利用统计学方法 AdaBoost[50,51]可以将各种链接预测算法进行线性组合，预测节点的链接情况，通过组合后形成的强分类器可以得到准确率较高的预测结果。但是该类实验仅仅是对预测算法的组合，没有给节点或边赋予不同的权重以适应不同的网络环境，本文在 AdaBoost 算法基础上，通过改变网络的节点权重，更精确的预测节点的链接预测。

1.3 本文工作

社交网络的迅速发展也引发了对社交网络的研究，将不同的社交网络抽象成网络结构图，从图形的结构去研究社交网络，是理论研究中经常会使用的方法。在由节点和边组成的网络图中，节点可以表示用户、蛋白质等单个关系实体，边表示节点与节点之间的联系，比如用户和用户之间的链接关系。不同于以往对社交网络大多侧重于从整体上进行研究，比如研究社交网络中所有节点或边的分布情况等，文章将社交网络的基本组成部分分解成节点和边，并从节点和边的角度单独的考虑，针对节点和边的不同特征，具体工作如下。

首先，阐述了目前对社交网络研究的工作。文中选取了对社交网络中节点和节点关系的研究方向的立意，并给出了以往在针对社交网络的研究中一般遵循的方向和研究思路。对社交网络的一般研究侧重于对其整体结构的研究，比如对社交网络中节点度、中心度等的研究，阐述了一般的社交网络的结构特征，但是对社交网络整体的研究太笼统，没有发现有关某特殊节点、特殊边的针对性。

其次，对社交网络中节点及其关系的研究进行了详细的阐述。对社交网络中节点研究中，每一个节点在社交网络中的作用和影响力不一样，如何度量节点的影响力，也成为社交网络中节点研究的重点之一。

在对社交网络中节点的研究中，每一个节点在特定的社交网络中具有不同的作用或地位，对目前影响力的估量方法进行了系统的整理、分类和分析，从各个不同的角度权衡影响力度量方法，并为本文的研究进行了铺垫。

本文主要考虑怎样评价社交网络中节点的影响力。文章所选的社交网络是微博。在微博社交网络中，每一个用户都有或大或小的影响力。影响力在不同的领域，具有不同的解读，文章认为，用户信息传播的越快，该用户具有较大的影响力。

在微博社交网络中，影响用户信息传播的因素有很多，比如用户的粉丝，一方面当用户粉丝较多时，用户所发的信息可以被更多的人看到，其信息也相应被更广泛的传播。另一方面，拥有较多粉丝的用户一般在网络中也被更多的人追捧，比如名人。粉丝作为一个度量指标，和用户的影响力呈正相关；同理，微博中用户的评论数和转发数也对其影响力有积极的作用。当用户所发的信息被更多的人评论和转发，其信息也被更快的传播；用户积极的在微博社交网络中发帖、转发等，与其所在的微博社交网络互动频繁，其信息也可以被更广泛的关注。

在对制约微博社交网络中用户影响力的诸多因素进行度量后，本文将各种度量指标分为两类，即用户活跃度和用户所发微博质量。结合度量指标，给每一个用户赋予一个权重，在微博社交网络图中，计算每一个用户的影响力。

另一方面，对目前社交网络中节点之间链接的相关研究进行了梳理。按照社交网络中节点边的大致研究方向，分为对节点属性的研究和网络拓扑结构的研究，对两者不足进行的改进。对边赋予不同意义的权重，在加权和不加权的社交网络中，分别预测社交

网络中节点对的链接关系。按与节点联系紧密程度分类，有强链接和弱链接的分类，研究了社交网络中弱链接的作用。结合目前大数据的发展趋势，改进链接预测的方法，使得链接预测能够应该在不同的应用领域中。

这些研究方法，都为链接预测算法提供了参考和借鉴。在对社交网络中边的研究中，本文主要侧重于社交网络中边的预测。通过链接预测可以发现网络中潜在的关系，可以扩展整个网络图。比如网站推荐系统的应用、基因之间作用所形成的网络图等。对链接预测的研究也主要分为对未来节点对的研究和对未知节点对的研究。

文章所针对的链接预测主要是对未来节点的链接预测，即已知在时间 t，网络节点的链接情况，预测在时间 t'（t'>t）的链接情况。选取的数据集为论文合著数据集，由监督学习的方法，在训练数据集中提取节点对链接的特征，在测试数据集中预测节点对的链接情况。

在链接预测问题中，计算每两个不同的节点之间的链接概率，概率较大其链接的几率也较大。在对链接预测的研究中，可以从两个不同的角度去考虑链接预测，即通过网络节点的邻居节点来考虑节点的属性，或者通过网络拓扑图，提取节点的网络结构特征。节点的链接与节点自身邻居数等属性有关，同时也与其所在的网络有密不可分的关系。本文为综合考虑链接预测的方法，采用了基于 AdaBoost 的方法，将每一种预测算法看作是一个弱分类器，将其线性组合成强分类器，进而预测节点对之间的链接情况。强分类器通过提高分类准确率较高的弱分类器所占的比例，综合考虑了弱分类器的不足之处，可以对链接预测有较好的效果。

最后，针对社交网络中节点及其关系的研究中，给出了最近的研究方向和新思路，新的研究也是未来继续学习和研究的方向。

1.4 本文结构

本文分 5 个章节来阐述社交网络中节点及其关系的研究，具体章节安排如下：

第一章绪论。主要阐述了研究背景、意义等，并详细介绍了目前的研究现状、本文的工作、文章的大体结构框架等。绪论主要用以帮助了解本文研究的理论和现实意义，系统的概括目前的研究情况和最新的研究进展，并给出文章的框架、大体脉络。

第二章相关技术与问题定义。提炼出文章用到的核心的相关算法和技术，包括算法的主要基本思想等，界定了实验部分不同的评价指标，并针对本文的具体算法，找到解决问题的方法。

第三章微博社交网络中用户影响力的评估。文章选定的社交网络是微博，在微博中评价不同的用户影响力情况。为了更真实客观的反映用户的微博影响力情况，本章节考虑到微博社交网络的特征，提出了新的影响力评估模型，在不同的评价体系中，衡量微博用户的影响力情况。

第四章基于 AdaBoost 提升算法的节点关系预测。主要是利用 AdaBoost 的算法，预测不同的社交网络中的链接情况。从链接预测的角度研究社交网络中节点之间的关系。

第五章结论。对本文的研究进行了归纳总结，并给出了下一步的研究工作。

2 相关技术与问题定义

2.1 相关技术

2.1.1 PageRank 算法

互联网的快速发展，带来了海量的数据和信息。网页承载的信息量大，且质量良莠不齐。如何在不断增长的海量信息中准确地定位到想要查找的信息是每一个人都很关注的问题。当网页能够按照某种重要程度降序地呈现出来时，可以把网页排在比较靠前的推荐给用户，那么用户就不需要在繁杂的网页中漫无目的地查找信息，可以极大的提高查找信息的效率，极大的方便人们的生活。

PageRank 算法正是目前应用较多的网页推荐算法。PageRank 算法是 Sergey Brin 和 Lawrence Page 于 1998 年提出了一种静态网页评级算法，主要对静态网页进行权威评判。对不同的网页进行 0-10 不等的评分，评分越高，网页也越具权威性。PageRank 算法通过网络中的超链接确定一个网页的等级评分，比如网页 X 链接到网页 Y，解释为网页 X 对网页 Y 的投票。根据投票的来源确定新网页等级。可以看作是一个网页权威值迭代的计算过程。对每一个网页赋予一个相同的初始值，网页最终权威值收敛到比较平稳的值，与初始值的选取无关。

PageRank 算法可以从两个方面来描述：网络数量和网页质量。从网页数量的角度，PageRank 算法中，被链接的网页越多，即网页的入度越多，则网页具有较高的权值；从网页质量的角度，一个网页如果被权威性较高的网页链接，则网页具有较高的权威。抽象出的网络图中，节点表示网页，边表示网页之间的链接关系。也可以根据链接关系的方向将社交网络分为有向的社交网络和无向的社交网络。有向的社交网络中，(i,j)表示节点 i 到节点 j 的链接，而无向的社交网络中，(i,j)表示节点 i 与节点 j 的链接，没有具体的方向性。常用的符号 G(V,E)表示社交网络，V 和 E 分别表示网络图中的节点集合和边的集合。

PageRank 算法可由以下公式表示。

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O(j)} \tag{2.1}$$

其中，P(i)和 P(j)分别表示网页 i 和网页 j 的权威值。E 表示网络图中的边的集合，(j,i)表示从网页 j 链接到网页 i，O(j)表示网页 j 的出度。

由以上公式也可以看到，当出现特殊情况的网页时，比如某个网页没有链出数，则网页的权威迭代不会收敛，在此算法的基础上，增加了阻尼系数 d，阻尼系数 d 表示用户继续点击进入网页链接的概率，同时用户有（1-d）的概率随机访问任意的网页。公式经过修正后，如下所示。

$$\mathrm{P(i)} = \mathrm{p} + (1 - \mathrm{p}) * \sum_{(j,i)\in E} \frac{P(j)}{O(j)} \tag{2.2}$$

阻尼系数的引入，可以使得网页链接中的权威值稳定的延续，不会出现中断或无限放大的现象，一般 p 的值取经验值 0.85。

PageRank 算法最大的优点就是可以防止作弊。网页的权威值比较高主要因为其拥有众多的网页链接，或者有重要的网页链接到该网页。一个网页很难将链入该网页的链接改变链接到其它网页，因此改变 PageRank 值并不容易。

因计算得到的 PageRank 值均与网络的拓扑结构有关，没有考虑查询的关键字，忽略了与某一个具体主题相关联的特征，导致按照搜索得到的网页排名与查询的关键字匹配降低。而静态的计算方法，也让 PageRank 算法厚此薄彼，偏向于旧网页。旧网页不断的积累，有更多的网页链接，得到的 PageRank 值也相应的较高，反之，新网页得到的值相对也较低，对新网页存在歧视，而实际上，可能新网页包含更多或更新的信息。通过 PageRank 值可以看出，最后是对每一个链入的网页的权威值进行了平均分配的加权，而事实上这样的“平均分配”有时并不是很符合实际的情况，每一个网页都偏向于链接到具有某一个特征的网页中，而不是随机的链接。

虽然提出的 PageRank 算法最初是针对网页评价，但是 PageRank 算法作为网络中随机游走算法的一种，被广泛的应用在各种不同的领域。能够抽象成节点和关系的网络图，节点可以看作是网页，节点之间的关系所形成的边可以看作是网页之间的投票行为。比如 PageRank 随机游走算法，可以发现网络的扩散情况，信息在网络的传播模型等。在微博社交网络中，节点表示微博用户，节点之间的链接关系表示用户之间的互粉。根据 PageRank 算法思想，如果每一个用户拥有较多的粉丝，或者拥有权威比较大的粉丝，则认为该用户具有较大的影响力，由此计算得到微博中用户的影响力。根据不同的影响力情况，有针对性的植入广告或推荐，或者找到在某一社交网络中的权威人士等。

文章针对 PageRank 算法存在的缺陷进行了改进。每一种社交网络都有其不同的特征，在微博社交网络中，微博中固有的转发、评论、发帖等，可以作为每一个节点特有的属性。每一个节点不同的属性特征也区别于其它的节点。具有不同属性的节点在网络图的随机游走中，也是以不等的概率将其权威值分配到其出度节点，并迭代计算每一个节点的影响力。

在利用 PageRank 算法思想时，给节点赋予权重，权重具体考虑微博社交网络中用户的评论、转发等行为，按照权重的比例不同，以随机游走的方法，计算每一个节点的影响力情况，按照值的大小排序，较大值的节点具有较大的权威值，其影响力也较大。

在计算微博社交网络的用户影响力时，微博用户之间的相互关系，概率链接等，将用户之间的关系看作是随机游走的网络关系，也可以利用 PageRank 算法的思想，计算每一个微博用户的影响力。

2.1.2 AdaBoost 算法

AdaBoost 是指提升算法，是统计学习方法中的一种。AdaBoost 算法主要是对若干个方法进行迭代的过程。通常被迭代的若干个方法叫“弱分类器”，由 AdaBoost 算法迭代后所形成的算法叫“强分类器”。

这样的迭代思想也可以简单的理解为“三个臭皮匠赛过诸葛亮”。多个专家综合判断比任意一个专家进行独断效果好很多。Kearns 和 Valiant 也从科研的角度给出理论支持。如果一个分类、预测等学习问题，有一个较高的准确率得到其学习的结果，那么这个学习问题就是一个“强可学习”问题，反之，如果学习得到的准确率较低，相当于随机概率，那么这个学习问题就是一个“弱可学习”问题。无论是强可学习还是弱可学习。Schapire 证明“强可学习”和“弱可学习”二者是等价的。一个学习问题是强可学习与这个问题是弱可学习是可以转换的。

找到准确率较高的强可学习方法并不容易，但是找到弱可学习的方法还是很容易的。于是，怎样通过弱可学习方法构造强可学习方法，也是提升算法需要解决的问题。本文应用到的主要是 AdaBoost 提升算法。

AdaBoost 算法主要在分类、预测等领域得到广泛的应用。其中预测问题可以看作是一个多分类问题，也可以由分类问题的思想来解决。将每一种分类算法看作是一个弱可学习方法，也叫“弱分类器”，而最后迭代而成的强可学习方法，也叫“强分类器”。通过弱分类器和强分类器对所给数据集进行分类。

在利用 AdaBoost 算法中，将弱分类器通过迭代组合的方式，合成强分类器，利用强分类器对数据集进行分类，可以得到更好的分类效果。在对数据进行分类，预测和学习时，每一个分类器都可以从不同的角度进行分类，但是每一个分类器所采取的侧重点也不尽相同，AdaBoost 算法通过整合弱分类器的分类优点，使其形成强分类器，提高分类学习的效果。

弱分类器如何组合成强分类器是 AdaBoost 算法的核心。当弱分类器进行分类时，提高分类错误的数据权重值，使分类错误的数据在下一个弱分类器中因为具有较大的权重值，而更容易被当前的弱分类器发现。不断改变数据集中数据的权重分别情况，当数据集被多个弱分类器学习后，分类错误的数据更容易被发现，所得到的强分类器分类错误的概率将减少很多。

强分类器是通过不断的修正弱分类器的数据权重分布，按比例的组合弱分类器而得得到的。将每一个弱分类器在数据集中进行学习后，会得到不同的准确率，为提高准确率较高的弱分类器所占的比例，将弱分类器的准确率作为其系数，将多个弱分类器进行系数比例的加权，最终得到强分类器。

AdaBoost 算法的大体流程如下所示。

(1) 首先对数据集中的每一个样本数据赋予一个相同的初始权重；

$$\mathrm{D} = (w_{11}, \mathrm{w}_{12}, \ldots, \mathrm{w}_{1N}),\ \mathrm{w}_{1\mathrm{i}} = \frac{1}{N}$$

其中，D 表示数据集，w_{mj} 表示在第 m 个弱分类器中的数据集中第 j 的数据的权值。

(2) 使用弱分类器对训练数据进行学习，得知每一个弱分类器分类的准确率 e；

$$\mathrm{e} = \sum_{i=1}^{N} w_{mj} I$$

其中 I 表示指示函数，当分类正确时，I 的值为 1，分类错误时，I 的值为 0.

(3) 由分类的准确率，增加分类错误的数据权重值，并更新数据集中数据样本的权值分布；

$$\mathrm{D} = \left(w_{m+1,1}, w_{m+1,2}, \ldots, w_{m+1,n}\right)$$

$$w_{m+1,j} = \frac{w_{mi}}{Z_m}\exp(\text{-eI}G_m(x)),\ \mathrm{i}=1,2,\cdots,\mathrm{N}$$

其中 Z_{m} 为规范化因子。

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-eIG_m(x))$$

其中 $\mathrm{G}_{\mathrm{m}}(\mathrm{x})$为第 m 个弱分类器。

(4) 计算每一个弱分类器的系数；

$$a_m = \frac{1}{2}\log\frac{1-e}{e}$$

(5) 由下一个弱分类器对数据集进行学习分类，重复(2)到(5)的运算；

(6) 最后由多个弱分类器，按照系数比例线性组合成一个强分类器。

$$\mathrm{f(x)} = \sum_{m=1}^{M} a_m G_m(x)$$

$$\mathrm{G(x)} = \mathrm{sign(f(x))}$$

最终得到的 AdaBoost 算法，对分类数据有较好的分类效果。整个 AdaBoost 算法思想和应用都便于理解和实现，不需要额外的特征选择，通过简单已有的弱分类器就可以得到一个强分类器。根据对 AdaBoost 算法的描述，可以得知，在每一个弱分类器迭代的过程中，都会对分类错误的数据加大权重，以便在下次分类时突显出来。但是如果该分类错误的数据在多个弱分类器中都被分类错误，那么该数据的权重将便会很大，对分类器准确率会有影响，甚至会左右弱分类器的选择。而往往这样的数据属于“边界数据”，处于分类正确与分类错误不确定最大的状态，同时也是对分类器影响最大的数据。如果训练数据集中存在噪音数据，对数据集中数据的权重分布也会有干扰，进而影响分类器的准确率。强分类器由多个不同的弱分类器组成，迭代训练时间较长，强分类器的分类效果也依赖于弱分类器的选择，弱分类器的个数，类别都会对强分类器的分类、预测效果、迭代时间等有影响。

AdaBoost 算法主要应用在分类、预测等领域。其中预测也可以看作是一个多类分类问题，在人脸识别等领域也有广泛的应用。本文主要利用 AdaBoost 原理实现社交网络中节点之间的链接预测。将节点之间的链接预测看作是一个二类分类问题，即链接或不链接。从不同的角度选择多个链接预测算法作为本文的弱分类器，由弱分类器组合成强分类器，预测节点之间的链接情况。

2.1.3 评价体系

为了说明提出的算法具有可行性、创新性或者在某一方面具有优越性，需要评价指标进行评估。对某一个具体的问题，可能也要从不同的角度按照不同的侧重点去评判，就需要有一系列针对该问题的评价体系。本文中的评价体系主要是指针对社交网络中节点及其关系的实验进行的评判标准。针对实验的结果展示，如何去评判，不同的实验有不同的评价指标。

在每一组实验中，会设定一个或多个对比实验。对比实验的设定可以作为新的实验的评价体系之一。将当前的实验和对比实验进行比较，分析不同的结果，可以得知当前对该类问题的解决情况和研究进展。作为对比实验，和当前的实验一定存在某些相似，或者是在对比实验的基础上进行改进，或者推进了对比实验的进度等，同时对比实验和当前实验都是在同一个数据集中的实验，对同一类问题进行解决，所以可以将对比实验看作是评价指标之一。

针对社交网络中节点的研究，从不同的角度进行评价。在研究微博社交网络中用户影响力的评价指标中，首先需要对影响力有一个界定，影响力的界定范围不同，对最终的结果也不一样。文中认为微博社交网络中，用户的信息传播的越快，用户的影响力也越大。最终得到的结果是影响力大小的降序排名，怎样评价这个排名的结果需要从不同的方面来考虑。

在研究微博社交网络中节点影响力的评估中，选取了多个评价指标作为文章的评价体系。首先是对比实验的选取，通过对某一算法的改进，在相同的数据集中进行实验，并对比最后的排名情况。其次选用了一般常用的针对排名的评价指标，具体包括 Spearman 相关系数，Kendall 相关系数，重叠率等，这些指标从不同的角度衡量两组排名的合理情况。另外，由于微博用户影响力排名的特殊性，文章采取了一种人工排名的方式作为对比实验之一。之所以这样实验，是因为微博用户的影响力排名实际上是一个主观的排名，从不同的角度，不同的用户，都会有不同的排名情况，没有一个统一的正确排序情况。虽然没有一个标准的排名情况，但是对某一问题的认识中，会有一个大致的排名，排名的偏差并不大。文章中采用了交叉验证的方法，多组实验数据交叉进行，最后采取平均值来作为衡量的标准，所得的实验结果更真实可靠。最后，文章比较了一下各个实验的计算代价，所耗的时间和内存空间，算法复杂度等。通过一系列的评价体

系，对最终得到的微博社交网络中用户影响力的排名有更客观准确的评价。

在对社交网络中节点关系的实验中，也需要从不同的角度评价实验结果。文章采用的是基于 AdaBoost 提升算法的链接预测算法，因为 AdaBoost 是在弱分类器的基础上进行，所以文章所选用的弱分类器将作为对比实验。评估多分类的链接预测问题，也可以从不同的角度去评价，所选的评价标准为准确率、召回率、ROC 曲线和 AUC 的值。

在链接预测的评价指标中，准确率主要用以评价预测的准确与否，直观的表现链接预测的结果。召回率是一种查全率。文中主要在交叉数据集中实验，从不同的方面去估量链接预测的结果。ROC 曲线主要是针对分类器的衡量指标，在针对分类器模型中，虽然准确率和召回率都相同，但是模型对分类错误和分类正确的惩罚不一样，通过 ROC 曲线反映分类错误和分类正确的比例情况，主要是从图形曲线的角度分析结果。而 AUC 将 ROC 的曲线面积转换成具体的数值，更直观的反映分类器的分类情况。

整体来说，对链接预测结果的衡量中，得到都是数值化的结果，可以很直观的得知基于 AdaBoost 算法得到的链接预测结果和其它的弱分类器得到的结果之间的差异或优势，说明基于 AdaBoost 算法的链接预测算法的可行性和合理性。

本文在社交网络中节点和节点关系的研究中，采用了不同的评价指标，从不同的角度衡量实验的结果情况，同时也考虑了每一个问题的特殊性，需要有不同的评价指标。通过对比实验和各种评价体系，说明本文所选取的算法具有合理性和可行性。

2.2 问题定义

本文所选取的研究内容是社交网络中节点和节点关系，针对这一问题，需要从不同的角度加以分解和定义。

在对社交网络中节点的研究中，首先需要在特定的社交网络中，确定单个节点的研究。每一个社交网络都具有不同的特点，将社交网络抽象成的网络图中，节点也代表不同的含义。节点可以具体的是一个用户、一个蛋白质基因等，节点是组成社交网络的基石，而每一个节点在社交网络中的作用、重要程度并不相同，怎样找到社交网络中比较重要的节点也是一个炙热的研究问题。在不同的社交网络中，“重要性”也具有不同的意义，可以指网络图中，用以链接不同的社交网络的关键节点，也可以指网络中连接边比较多的活跃节点，或者是一个社区中的领袖人物等。这些节点都被认为是在某一个社交网络中比较特殊、比较重要的节点。找到这样的节点，对社交网络的扩充，社交网络中节点推荐等都具有研究意义。

文章选择的社交网络是微博社交网络，节点即为每一个微博用户，主要研究微博社交网络中用户影响力评估。微博社交网络具有其自身的特定和属性，每一个微博用户在其平台中转发、评论微博，具有特定的用户行为，同时，在微博中，某一用户具有粉丝，被评论、被转发等社交网络中其它用户的行为，可以从主动和被动两方面来衡量微博用

户的影响力情况。

在对“影响力”的界定中，文章认为微博用户信息传播的越快，其影响力也越大。在微博中，每一个用户所发的信息被更快的广为人知，该用户的影响力就更大。在衡量微博中的用户影响力时，充分考虑到微博的特征，将用户的主动行为，包括用户所发微博，用户转发、评论等互动行为，结合用户的被动行为，即用户所发微博质量，包括用户微博被转发、被评论的次数等，得到每一个用户的影响力权重，最后在整个微博社交网络中利用所得到的权重，衡量用户的影响力。

在研究社交网络中节点关系时，文章主要是有关社交网络中节点连接与否的研究。在对节点的链接中，在合著论文的数据集中，预测节点对的链接情况。

本文对节点链接预测的研究中，将数据集分为训练集和测试集，在训练数据集中找到与节点链接相关联的属性，包括节点自身属性和节点所在网络的网络拓扑结构等，用以预测在测试集中节点的链接情况。将数据集按多倍交叉验证的方法，多次训练，最终得到预测结果的平均值。

在具体的链接预测方法中，将链接预测问题看作是一个多类分类问题，采用的是基于 AdaBoost 的提升算法，将一般的链接预测算法看作是一个对比实验，也是所选取的弱分类器，将得到的最终结果作为一个强分类器，比较不同的结果。

对不同研究内容，将从不同的角度采用多个评价指标，即评价体系，对结果进行综合的评判。评价体系从数据的标准说明了提出问题的优越性分析，让实验的结果更有说服力。评价体系的选取也依赖于对特定问题的研究，有针对这一类研究问题的评价体系，也有对研究问题提出了特殊的评价体系。比如，在对微博社交网络用户影响力分析中，就提出了一个人工排序的方法，更加客观合理的说明最终的排序情况。

在后续的章节中有对每一个问题细致的说明，将对社交网络中节点及其关系的研究分解，逐步的论述研究的不同方面，更加清晰的认识到对这一类问题的分析和研究。

2.3 本章总结

本章节主要对文中涉及的相关技术进行了归纳和总结，主要包括三个方面：PageRank 算法，AdaBoost 算法和评价体系。

在对 PageRank 算法的相关技术中，不仅仅说明了所提及的 PageRank 算法核心思想，该算法的来源，优点和缺点等，也说明了对该算法的进一步改进。所给出的 PageRank 算法只是对一个问题研究的核心思想部分，该算法中随机游走的思想适应于在其它社交网络中运用。

在对社交网络中节点影响力的研究中，主要运用了 PageRank 相关技术。同时考虑到微博平台的特殊性，将微博中用户的评论、转发等因素，刻画成具体的数据，更具有针对性，也对 PageRank 算法的不足之处进行了改进。

另一个相关技术主要是 AdaBoost 提升算法，文中详细的说明了 AdaBoost 的算法来源，实现过程等。在对社交网络中节点关系的研究中，主要是研究节点间的链接与否，采用了基于 AdaBoost 算法的相关技术，将链接预测问题看作是一个多分类问题。将其它分类器看作是一个弱分类器，改变数据集的分布，最后组合成一个强分类器，预测社交网络中节点的链接情况。

最后，还给定了针对结果的评价体系。评价体系主要是用以说明实验结果的真实可信。对每一类问题的研究中，通常都会有既定的评价指标，不同的研究问题需要采用不同的评价指标来评判，也可以给出对某一特定问题的评价情况。本章节给出了针对社交网络中节点影响力、节点链接关系的不同评价指标。

针对社交网络中节点影响力的评价体系中，有一般常有的评价指标，也提出了特定的针对微博社交网络中评价指标，即人工排名。在不同的评价体系中，衡量实验结果的合理性，说明排名的客观真实性。

在对社交网络中节点关系的评价体系中，所用到的评价体系多为一般常用的评价体系，将不同的排名情况从不同的角度进行比较说明，评价指标所得到的是将两组排名数值化的情况，更直观准确的了解链接预测的结果。

将相关技术和问题定义直接罗列出来，可以将问题进行细致的分解，提取不同问题所涉及的技术领域，对相关技术的改进。在对后续研究中所提及的技术问题，术语等有直观的了解和认识。

3 基于 PageRank 算法的社交网络节点影响力评估

3.1 问题引出

社交网络是由不同的参与者自发而形成的虚拟社交网络，每一个社交网络的参与者也扮演着不同的角色。研究社交网络中每一个节点的影响力在研究领域和应用领域均有非常重要的作用。发现网络中的关键节点，可以帮助构建和扩充社交网络图；商家对一个社交网络中最有影响力的客户进行广告植入，以最小的投资获得最大的利益。如何评价一个社交网络中节点的影响力也是本章节的研究目的。

本章节主要研究在微博社交网络中如何评价用户的影响力。在微博时代，每个人好像都有一个麦克，在 140 字的信息中分享着自己心情和故事同时表达了自己对世界的认识、参与热门微博的讨论，微博也成为各种观点和舆论的重要发源地，对信息的传播具有重要作用，同时微博具有评论、转发、关注等多种功能，可以加快信息的传播。每一个用户所发的微博都可以引起其他用户的关注、转发和评论，从而在微博社交网络中产生一定的影响，当然每一个用户的影响力也是截然不同的。

在制约微博用户影响力的众多因素中，文中认为用户的传播能力强，用户的信息可以更快的在网络中扩散，其影响力也越大。和传统的用户影响力的评价方法相比，文中综合考虑用户的活跃度和用户所发微博质量两个方面的因素，得到用户的影响力权重，然后把用户看作社交网络中的节点，计算节点在社交网络中的影响力。通过在公开语料集和真实数据中的实验，表明该方法是可行的，并比传统的用户影响力的评价方法更能客观真实的反映用户的实际影响力。

3.2 算法设计

3.2.1 加入权重的用户影响力评价方法

提出的加入用户权重的影响力评价方法主要考虑两个方面的因素：用户的积极活跃度和用户发布的微博质量。其中用户的积极活跃度包括用户发布微博、转发微博的频率，活跃度反映了用户参与微博互动的热情和积极性，用户的活跃度越高说明了其与其他博主的互动频繁，更新微博较快，更新的微博的信息可以引起其他用户的好奇而关注围观，进而加快该博主的信息传播。用户发布的微博质量包括用户微博被转发和被评论的次数，用户所发的微博质量越高，越容易引起大众的转发和评论，该博主的微博信息也被传播的更快，同时该微博用户也更引起众人的关注。无论是用户活跃度还是用户发布微博的质量，都是和用户的影响力呈正相关。如下图 3.1 所示。

用户的权重计算公式如下所示：

$$W_i = X_i + Y_i \tag{3.1}$$

X_i是指用户 i 的活跃度。计算用户 i 的活跃度X_i时考虑用户在某一时间段内原创微博的数量P_i和转发微博数量R_i,具体的计算公式如下：

$$X_i = \frac{P_i}{T} + \frac{R_i}{T} \tag{3.2}$$

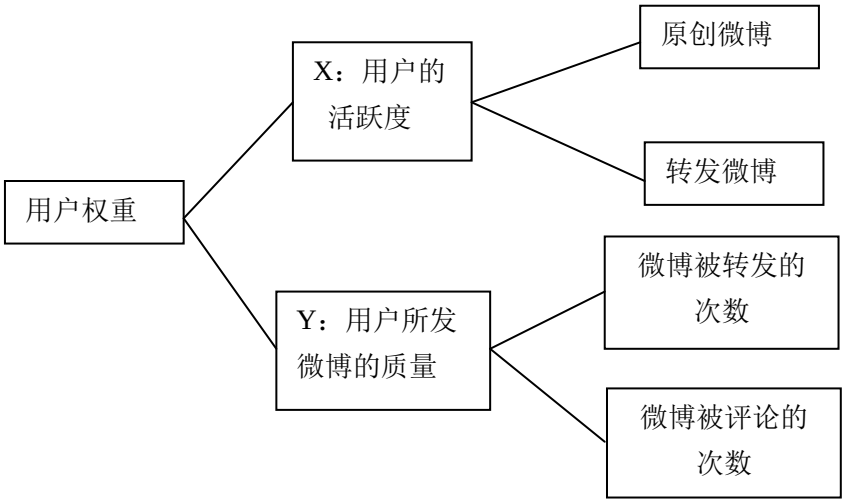

图 3.1 用户的权重及相关因素

Fig.3.1 The weights of users and the related factors

Y_i是指用户 i 发布的微博质量。Y_i是用户在一段时间 T 内所有微博影响力的平均值，评价每一条微博的影响力主要考虑微博被评论次数和微博被转发次数两个方面的因素，计算用户发布的每一条微博的质量计算公式如下：

$$y_{i,j} = \sqrt[3]{MR_{i,j}} + \sqrt{MC_{i,j}} \tag{3.3}$$

其中 $y_{i,j}$ 表示用户 i 发布的第 j 条微博的质量，$MR_{i,j}$ 表示用户 i 发布的第 j 条微博被转发的次数，$MC_{i,j}$表示用户 i 发布的第 j 条微博被评论的次数。得到用户 i 发布微博的平均质量计算公式如下：

$$Y_i = \frac{\sum_{j=1}^{n} y_{i,j}}{n} \tag{3.4}$$

其中 n 是指用户在某时间段内所发微博的数量，包括原创微博和转发微博。

3.2.2 用户排名的影响力评价指标

在对用户影响力排名评价时，常用的评价指标有排名的 Spearman 序列相关系数，Kendall 序列相关系数，重叠率，计算代价等。

其中 Spearman 相关系数反映的两组排名之间的线性相关，该值越接近+1 或-1，两组排名之间呈线性相关，Spearman 相关系数的符号反映了两组排名之间正相关和负相关的关系。符号为正号，两组排名呈正相关，符号为负，两组排名呈负相关。具体 Spearman

相关系数排名评价方法如下公式 3.5，其中 xi 和 yi 分别表示在两组排名中的排名序号，N 表示排名总数。

$$p = 1 - \frac{6\sum (xi - yi)^2}{N^3 - N} \tag{3.5}$$

Kendall 相关系数反映了 一组排名相对于另一组排名的分歧。其中如果两组排名是完全吻合的，该值为+1；如果两组排名分歧最大，该值为-1；两组排名的一致性和其值呈正相关。具体 Kendall 相关系数排名方法如下公式 3.6，其中 p 表示两组排名一致的对数，n 表示排名总数。

$$\tau = \frac{4\rho}{\mathrm{n}(n-1)} - 1 \tag{3.6}$$

重叠率 overLap 主要是指两组排名在前 N 名重叠的次数（记为 top N），N 的取值可以根据具体的数据集合理的选择。具体 overLap 如下公式 3.7，其中 topN(x)和 topN(y)分别表示两组排名中 topN 的次数。

$$overLap(x, y) = |topN(x) \cap topN(y)| \tag{3.7}$$

计算代价是指根据该方法得到的用户影响力评估所消耗的时间，空间复杂程度，以便于在具体的研究中根据应用和环境选择理性的评估方法。

3.2.3 对比实验描述

文中所选取的对比实验是用户的粉丝数、转发数排名以及参考文献[29]在网络拓扑结构中得到每一个用户的领袖排名即 leaderRank，其中领袖是指有影响力的用户。

按照用户的粉丝数排名。虽然粉丝数多的用户，其影响力不一定大，但是粉丝数也反映了用户的人气和流行度，该微博用户被很多的粉丝跟随，微博动态为更多用户所知，相应的信息也得到更广的传播。

对用户的转发数排名，用户的转发数越多，其信息可以更快的在网络中扩散，可以在一定程度上反映用户的影响力。

文献[29]是在网络拓扑结构的角度，改进了传统的 PageRank 算法而得到的领袖排名算法 leaderRank.在微博社交网络中，将微博用户看作节点，用户之间的“互粉”形成一条边，如下图 3.2 所示，节点 A 和 B 分别表示用户，A 到 B 的边表示 A 是 B 的粉丝。

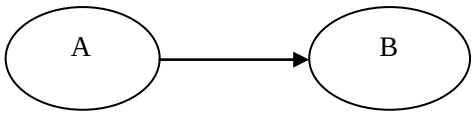

图 3.2 用户之间的关系

Fig.3.2 The relationship between users

LeaderRank 算法提出了一个虚拟的根节点（ground Node）,该根节点和网络中的节点形成双向的链接，如图 3.3 所示，其中实线的单向边表示节点之间的粉丝跟随，虚线

所形成的双向边表示根节点和网络中所有节点的链接。

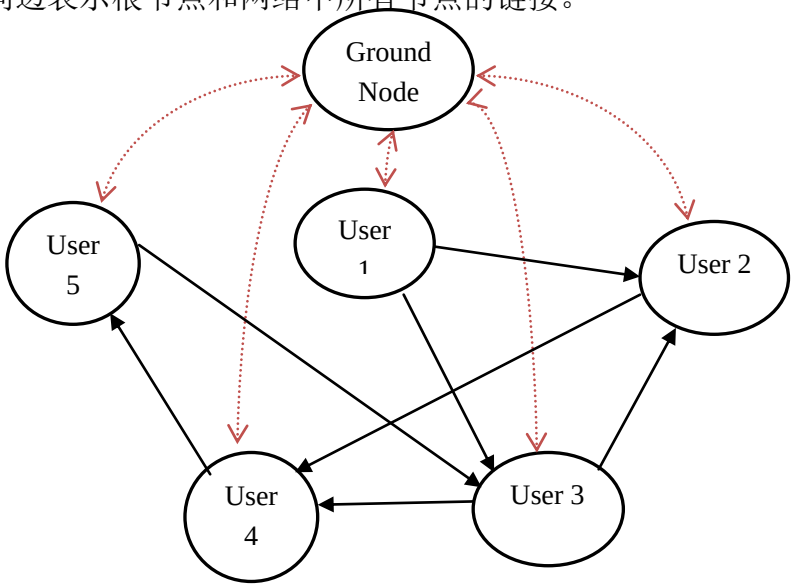

图 3.3 加入 ground node 节点后的网络图

Fig.3.3 The network graph after adding ground node

计算用户节点的影响力如公式 3.8 所示。其中 u, v 表示网络中的节点，M(u)是指向节点 u 的所有节点集合，N(v)是节点 v 的出度。从公式(8)可以看出计算每一个节点的 leadRank 值的过程是一个迭代的过程，其中初始化网络中每一个节点的 leaderRank 值(LR)为 1.0，而根节点的初始值为 0.0.

$$\mathrm{LR(u)} = \sum_{v \in M(u)} \frac{LR(v)}{N(v)} \tag{3.8}$$

得到每一个节点的 leaderRank 值后，和根节点的 leaderRank 值归一化进行累加，得到最终的 leaderRank 值，如公式 3.9 所示，排序，得到影响力较高的用户。其中 N 是根节点的出度，即网络中的所有节点数。

$$LR(u) = LR(u) + \frac{LR(g)}{N} \tag{3.9}$$

3.3 实验结果与分析

3.3.1 语料简介

本实验包括两个语料集，第一个是在公开的数据集上的实验。第二个是在真实的数据集中。

在公开数据集中本文选取的是 2012 年 kddcup track1 上的数据集，该数据集是腾讯微博爬起的共 90 天内的相关数据，将得到的数据从以下两个方面概括：

(1)用户属性：用户发布的原创微博数；用户转发、评论微博数；用户所发布的微博被转发、被评论的次数等。

(2)用户关系：用户之间的关注，用户的粉丝。

在真实数据集中，本文选取的是 2012 年新浪微博名人影响力榜 9 月份的数据。其中名人堂中的数据基本上是经过认证，数据信息比较真实可信，数据主要是用户的属性，包括用户所发微博，转发微博，用户在给定时间段内微博被转发、评论的次数等。

3.3.2 实验流程

本文选取了两个数据集，在公开数据集和在真实数据集中，针对不同的数据集，为了说明本章节所提出的加入用户权重的方法，和各个影响力排名进行了比较。根据数据集不同的特点，实验也有所区别。

在公开数据集 kddcup track1 的实验中，分别对用户的粉丝数，用户转发数，基于 leaderRank 算法的用户影响力排名和加入用户权重的影响力方法排名，比较在不同的评价指标中各个排名的异同。其中 Spearman 相关系数分别比较两组排名占总排名的 1%,10%和总排名的相关性。重叠率比较 top 10，top 20，top 50，top 100 出现相同节点的个数。

在真实数据集新浪名人微博社交网络中，分别对用户的粉丝数，用户转发数以及加入权重的影响力方法排名。其中 Spearman 系数比较占总排名 10%和总排名的线性相关性。因为在名人排名榜中，占总排名 1%的排名太少，所以没有比较 top 1%的 Spearman 系数。重叠率比较 top 10，top 20，top 50 的出现重复用户，真实的名人微博排行榜中数据集较小，所以没有比较 top 100 出现重复的用户个数。在该数据集中的实验没有基于 leaderRank 算法的用户影响力排名，因为在特定的真实数据集中，给定了用户，用户之间的相互关注链接太少，所得到的节点之间的网络拓扑图也很稀疏，网络结构信息太少，故没有从网络拓扑方面比较各个排名的异同。

3.3.3 实验结果及分析

分别在公开的语料集和真实数据集中实验，可以得到每种影响力排名下的前 10 名用户排名，从不同的影响力评价方法中，比较任意两组排名。具体的实验结果如下表所示。

表 3.1 在 kddcup track1 中各个算法的 Spearman 相关系数

Tab.3.1 Correlation coefficient of Spearman in kddcup track1 of each algorithm

Correlation	Top 1%	Top 10%	All
粉丝数 Vs 转发数	0.1365	0.1926	0.5718
粉丝数 Vs 基于 LeaderRank 算法	0.9030	0.7924	0.8600
粉丝数 Vs 加入用户权重	0.1024	0.4204	0.4395
转发数 Vs 基于 LeaderRank 算法	0.9677	0.7694	0.3378
转发数 Vs 加入用户权重	0.4256	0.9892	0.9624
基于 LeaderRank 算法 Vs 加入用户权重	0.6978	0.7558	0.2976

表 3.2 在 kddcup track1 中各个算法的 Kendall 系数

Tab.3.2 Coefficient of Kendall in kddcup track1 of each algorithm

	Kendall
粉丝数 Vs 基于 LeaderRank 算法	0.6788
粉丝数 Vs 加入用户权重	0.2673
转发数 Vs 基于 LeaderRank 算法	0.0665
转发数 Vs 加入用户权重	0.3166

表 3.3 在 kddcup track1 中各个算法的重叠数 overLap

Tab.3.3 Overlap in kddcup track1 of each algorithm

	Top 10	Top 20	Top 50	Top 100
粉丝数 Vs 转发数	1	2	4	10
粉丝数 Vs 基于 LeaderRank 算法	4	13	34	43
粉丝数 Vs 加入用户权重	2	3	8	17
转发数 Vs 基于 LeaderRank 算法	0	1	6	35
转发数 Vs 加入用户权重	1	7	20	52
基于 LeaderRank 算法 Vs 加入用户权重	1	2	6	36

表 3.4 在新浪微博名人排名榜中各个算法的 Spearman 系数

Tab.3.4 Coefficient of Spearman of each algorithm in Sina microblog celebrity rankings

Correlation	All	Top 10%
粉丝数 Vs 转发数	0.5421	0.5394
粉丝数 Vs 加入用户权重	0.5145	0.6424
转发数 Vs 加入用户权重	0.9468	0.5636

表 3.5 在新浪微博名人排名榜中各个算法的 Kendall 系数

Tab.3.5 Coefficient of Kendall of each algorithm in Sina microblog celebrity rankings

	Kendall
粉丝数 Vs 加入用户权重	0.3741
转发数 Vs 加入用户权重	0.8230

表 3.6 在新浪微博名人排名榜中各个算法的重复数 overLap

Tab.3.6 Overlap of each algorithm in Sina microblog celebrity rankings

	Top 10	Top 20	Top 50
粉丝数 Vs 转发数	3	5	31
粉丝数 Vs 加入用户权重	3	7	31
转发数 Vs 加入用户权重	7	14	48

表 3.7 在 kddcup track1 各个算法得到的 Top10 用户排名

Tab.3.7 The top 10 users of each algorithm in kddcup track1

排名	粉丝数排名	转发数排名	基于 LeaderRank 算法的排名	加入用户权重的排名
1	1760378	2121077	1774462	2012081
2	1774717	377344	330571	2105405
3	1760642	1579498	1774457	1185040
4	1760350	1100559	1774520	1412782
5	647356	1181393	1902321	2122259
6	1419930	463485	1774705	741560
7	1837210	1118081	1760654	1774586
8	1774505	866409	346092	1760310
9	1925678	830738	1954203	1341177
10	682877	414978	1774687	1774815

表 3.8 在新浪微博名人排名榜中各个算法得到的 Top10 用户排名

Tab.3.8 The top 10 users of each algorithm in Sina microblog celebrity rankings

排名	粉丝数排名	转发数排名	加入用户权重的排名
1	姚晨	加措活佛-慈爱基金	阿信
2	小 S	阿信	刘忻
3	谢娜	吴奇隆	王力宏
4	王力宏	刘忻	加措活佛-慈爱基金
5	何炅	李承鹏	吴奇隆
6	赵薇	张杰	蔡康永
7	杨幂	蔡康永	张歆艺
8	陈坤	张小娴	张杰
9	蔡康永	谢娜	小 S
10	林心如	王力宏	郭德纲

由表 3.1,表 3.2，表 3.4，表 3.5 在不同的数据集中各个排名评价指标中可以看出不同的影响力排名符号都为正号，都是正相关的，这说明了虽然影响力的评价指标不同，但是各个方法之间也不是毫无关联的，比如粉丝数多并不一定影响力大，但是粉丝数多其影响力不会太小。表 3.1 中粉丝数和转发数相关性较低，粉丝多的其转发不一定高；粉丝数和 leaderRank 算法相关度较高，根据粉丝数得到的影响力较高的，leaderRank 算法得到的影响力也较高，因为 leaderRank 算法是对 PageRank 算法的改进，而 PageRank 算法是与节点的入度（即粉丝数）相关的；转发数和加入用户权重排名相关性较大，两者都反映了用户的信息在微博中传播的速度，而这也说明了加入用户权重的影响力评价方法可以加快微博信息的传播。

表 3.2 也说明了基于 leaderRank 算法的影响力评价方法和粉丝数的评价方法更一致，而加入用户权重的评价方法和转发数的评价方法更一致。

从表 3.3 和表 3.6 的各个排名重叠数可以看出虽然各种排名有所区别，但是无论哪一种排名，随着 top N 中 N 的增大，重叠数也逐渐增大，各种排名之间有一定的相互联系。同时从表 3.3 中可以看出各个排名侧重点有所不同，排名很少重回。

表 3.4 和表 3.5，在真实数据集中，可以看出加入用户权重的影响力评价方法和用户的转发数相关性较大，用户的权重越大，用户的转发数越多，在微博中也传播的更快。而 信息更快更广的传播，用户实际的影响力也越大。

表 3.7 和表 3.8 是在不同的评价方法中得到的 top10 的用户排名,每种排名算法得到的排名侧重点不同，排名也有差异。由表 3.8 可以看出粉丝数较多的用户一般是活跃流行的明星，转发数较多的一般是微博所发的质量较高引起共鸣和关注的用户，比如知名媒体等。而加入用户权重的用户影响力排名综合考虑了各种因素，将用户的活跃度和用户所发微博的质量两个方面的因素结合起来，用户越活跃，所发的微博被转发和评论的次数较高，该用户的信息也可以更快的在社交网络中传播，更能为人所知，

在计算代价上，粉丝数影响力排名最简单直接，只需要一遍统计每一个节点的入度即可。转发数数需要累加用户在某时间内每条微博的转发数。leaderRank 算法则是基于网络拓扑结构计算用户的 leaderRank，得到节点的入度、出度等，还需要迭代的过程，相比较而言计算成本较大。加入用户权重的评价方法，需要得知或爬取用户发布的微博数和转发数，用户发布的微博被转发和被评论的次数，计算代价适中。

3.4 本章总结

本文针对微博社交网络中用户的影响力评估进行了研究，对微博用户的影响力界定为：微博用户的信息可以更快的在网络中传播，其影响力也越大。通过分析得到一种加入用户权重的用户影响力算法，该方法考虑了用户本身的活跃积极性和所发微博的质量两个方面。

比较传统的对用户属性的评价方法，包括用户粉丝数，用户转发数等，可以使信息在网络中得到更快的传播，在基于网络拓扑机构对用户影响力的评价方法中，本文通过给用户加入权重，计算其在微博社交网络中的权威，即影响力，可以突出用户的活跃度和所发微博质量两方面的因素，更加合理的反映用户的实际影响力。

通过本章节的研究，可以知道用户影响力包括用户的主动行为和用户所发微博被转发评论等被动行为（用户所发微博质量）。在实际的微博社交网络中，为了提高用户的影响力，用户可以通过活跃的参与微博的发帖、评论、转发、加强和其他用户的互动，这样可以引起更多人的关注，增加更多的粉丝，同时可以发布更多高质量的微博，引发大家对微博的转发、评价，引起更多的围观，让信息在网络中更快的传播，提高用户在微博社交网络中的影响力。

评价微博用户的影响力可以从不同的角度给出不同的影响力排名，每一种影响力因素的侧重点有所不同，怎样给出合理的，大众接受的影响力排名需要在不用的应用环境中区分考虑。下一步的工作可以从实际微博社交网络中影响微博用户排名的因素，比如微博认证，微博标签等方面来研究微博用户的影响力评价方法。

4 基于 AdaBoost 算法的社交网络节点关系预测

4.1 问题引出

社交网络中的每一个节点都不是孤立的存在，节点和节点之间都有相互的关系。比如在微博社交网络中，用户之间的互粉、关注等关系；在论文合著网络中，作者之间的合著关系；在生物领域，蛋白质之间的链接关系等。在已知的社交网络关系中，通过监督学习的方法，得知社交网络中，已经链接的节点对之间所遵循的属性或特征，从而预测在未来的时间内可能发生的链接或者当前未显示出来的链接对。

随着在线社交媒体的发展，各种形式的社交网络也是人们生活不可或缺的一部分，例如 Facabook，Twitter，微博，QQ 等。个体以及个体之间的各种关系就形成了一个无形的社交网络，每个个体都不经意的生活在一个虚拟的社交网络范围内。人们通过社交网络和朋友们保持联系，同时也可以找到和自己志趣相投的新朋友或者搜到自己想要的各种物品，而商家也可以预测潜在的客户群，针对性的宣传或进行精细化的推荐，得到最大的投资回报。而在大数据的时代，预测或者推荐能够让我们对一个人或者物体的描述全部数字化，通过数据或者行为方式等所抽象成的网络拓扑图可以对每一个个体形成一个数字画像，进而得到个体的行为模式，更精准的预测个体行为。

如何预测社交网络中的未知链接，本章节将从节点属性和网络拓扑结构两个方面进行研究。在对以往传统方法基于节点邻近度和基于网络拓扑结构的方法总结后，提出了基于 AdaBoost 提升算法对未知链接进行预测，给节点赋予权重，通过提高被前一轮弱分类器错误分类的节点权重，同时提高弱分类器的比例，将多个弱分类器线性组合成强分类器，使之预测节点之间的未知链接。实验在公开的数据集科学论文合著网络上进行，和传统的基于节点属性和基于路径的方法等单一的弱分类器都进行了比较，证明基于 AdaBoost 方法的强分类器能更准确的提高未知链接的准确率，提高分类性能。

4.2 算法设计

4.2.1 AdaBoost 算法

AdaBoost 算法属于一种统计学方法，主要是将若干个弱分类器组合成强分类器的过程。学习得到强分类器可以看作是一个“分而治之”的思想，通过一步步的得到弱分类器，将其组合成强分类器，提高分类，预测，概率计算等的效果。

在具体学习强分类器的过程中，主要分为两个步骤。一方面改变训练数据集中的权重，另一方面线性组合弱分类器。改变训练数据集中的权重主要是指提高在这一轮中分类错误的样本权重，这样该样本可以被下一个弱分类器关注，被正确分类的概率相应的也提高了。线性组合弱分类器主要是指提高分类准确率较好的弱分类器比重，使得分类

准确率较好的弱分类器在最终的强分类器中可以起到较大的作用。

本章节基于 AdaBoost 算法思想，将节点对的预测看作是一个二类分类问题，即节点对有链接或没有链接两种情况。弱分类器和强分类器主要是指分类预测的算法。得到弱分类器较容易，通过给样本赋予权重，提高在这轮预测算法中分类错误的样本权重，使其更好的被下轮的预测算法关注，并为分类准确率较高的弱分类器赋予较大的比重，使得多个预测算法在组合成强分类器时，分类效果较好的弱分类器具有较大的作用。组合各种不同的弱分类器形成一个新的链接预测算法也相当于综合考虑了不同的链接预测算法，从不同的角度来考虑影响链接预测的因素。

4.2.2 弱分类器的选择

文中的弱分类器是指每一种链接预测的算法。链接预测算法有基于邻近度方面的算法和基于中心度方面的算法。

(1) 基于邻近度方面的算法

基于邻近度方面的算法主要是利用节点的邻居节点信息得到的一类算法。基于邻近度主要是针对节点的属性来说，通过提取网络中节点的邻近节点，节点度等属性特征预测节点对之间的链接情况。在这类算法中，共同邻居节点成为主要的考虑因素，在这一基础上提出和改进了其他的预测算法。

比较节点 i 和节点 j 之间的邻居节点。有以下基于邻近度方面的链接预测算法。

共同邻居节点数。节点 i 和节点 j 中之间的邻居数，公式如下：

$$f_{cos(i,\ j)} = |\{k|i\sim k\ \wedge k\sim j\}| \tag{4.1}$$

其中节点 k 表示和节点 i，j 都有连边，即节点 k 是 i 和 j 的共同邻居。

Jaccard 系数通过改进共同邻居节点预测方法，公式如下：

$$f_{jac(i,\ j)} = \frac{|\{k|k\sim i\wedge k\sim j\}|}{|\{k|k\sim i\vee k\sim j\}|} \tag{4.2}$$

Adamic/Adar 方法对共同邻居数进行了计数，同时也考虑了邻居节点的度，d(k)表示节点 k 的度。公式如下所示：

$$f_{AA(i,\ j)} = \sum_{k\sim i\wedge k\sim j}\frac{1}{\log(d(k))} \tag{4.3}$$

最后两个基于邻近度的链接预测算法主要是基于图的邻接矩阵。由节点和连边构成的网络图 G(V, E)，其中 V 表示网络节点集合，E 表示节点之间的边的集合。A 是一个 n*n 的矩阵，表示图的邻接矩阵，显然在无向的网络图中，邻接矩阵 A 是一个对称矩阵。如果节点 i 和节点 j 有连边，A_{ij}=1，否则 A_{ij}=0.

基于邻近度的指数函数链接算法公式如下：

$$f_{exp(i,\ j)} = [e^{aA}]_{ij} = \sum_{p\in P*(i,j)}\frac{\alpha^{|p|}}{|p|!} \tag{4.4}$$

其中 a 的系数一般取经验值0.85.

基于邻近度的诺依曼图形核函数的链接算法公式入下所示：

$$f_{NEU(i,j)} = [(I-aA)^{-1}] = \sum_{p \in P*(i,j)} \alpha^{|p||} \quad (4.5)$$

其中 P*(i, j)表示节点 i 到节点 j 的所有路径。

(2) 基于中心度方面的算法

基于中心度的链接预测算法主要是指基于网络拓扑结构，考虑节点在整个网络图中的偏好链接。该类方法主要是指两个节点 i 和 j 的中心度乘积，选择不同的中心度度量方法也形成了不同的链接预测算法。其中主要的两种度量方法如下所示：

$$f_{PR(i,j)} = PR(i)*PR(j) \quad (4.6)$$

其中 PR(i)和 PR(j)指基于 PageRank 算法得到的节点 i 和节点 j 的中心度的值。PageRank 算法主要用来对网页进行评级。在社交网络中，也可以将一个节点看作一个网页，利用 PageRank 算法计算每一个节点的 PR 值。

$$f_{PA(i,j)} = d(i)*d(j) \quad (4.7)$$

其中 d(i)和 d(j)分别指节点 i 和节点 j 的度。

4.2.3 强分类器的形成

强分类器是以上弱分类器的加权组合。给每一个节点赋予一个权重，当弱分类器对其进行分类时，如果得到的链接预测错误，则提高相应的节点权重，以便下一个弱分类器发现这个节点，提高该节点被正确分类的概率。同时由每一个弱分类器的精确率得到该分类器线性组合的系数，这样精确率高的分类器可以占较大的比例。对分类错误的节点对，通过增大权值，在以后的分类中更容易被预测到，组合而成的强分类器准确率也较高。

理论上来说，以上每个弱分类器都要参与到强分类器的线性组合中。本文选取了所有的弱分类器进行线性组合，同时也根据实验的结果，选取了两类链接预测方法中效果最好的两种弱分类器，只组合这两个最好的弱分类器，比较和分析不同的弱分类器对链接预测的影响。具体算法如下所示。

(1) 初始化训练集中每个样本节点的权重。在本文中初始化的节点权重 w 均为1.0.

(2) 根据 AdaBoost 算法反复学习弱分类器。计算在每一个弱分类器 G(x)中，链接预测的准确率 e，计算每一个弱分类器的系数 a，其中 a 的计算如下所示：

$$a = \frac{1}{2}\log\frac{1-e}{e} \quad (4.8)$$

(3) 更新每一个节点的权重，w←w+e

(4) 在新的节点权重形成的网络结构中，利用下一个弱分类器进行分类预测。

(5) 最终将多个弱分类器加权组合成强分类器，其线性组合得到的强分类器如下。

$$G(x) = sign(f(x)) = sign(\sum_{m=1}^{M} a_m G_m(x)) \quad (4.9)$$

其中 a_m 是弱分类器的系数，$G_m(x)$是指弱分类器。m 是指分类器的个数，在本文中 m 取 7 表示弱分类器由 4.1-4.7 中的弱分类器组成，m 取 2 表示在两类分类器中选取分类效果最好的两个弱分类器。

4.2.4 对比实验描述

本章节的对比实验包括选取的所有弱分类器，即公式4.1-4.7中的链接预测算法。比较在不同的数据集中弱分类器和得到的两个强分类器之间链接预测的差异和优劣。

单一的弱分类器之间的相互比较，可以得到这两类算法中，对链接预测效果较好的预测算法，基于节点属性和基于网络路径，对链接预测影响力较大的方面；和形成的两个强分类器进行比较，可以得知弱分类器和强分类器之间的链接预测效果；比较两个强分类器的分类效果，可以得知弱分类器的组合数对强分类器的影响。

4.3 实验结果与分析

4.3.1 语料简介

文中所选用的实验数据是公开的论文合著数据集，即 arXiv 数据集。该数据集包括不同方向的论文合著数据集，分别有 CA-hepth, CA-astro, CA-condmat, CA-grqc, CA-hepph 等5个不同的数据集。数据集中包括论文的作者以及作者之间的合著情况。将每一个作者看作是一个节点，合著的作者之间看作是一条连边，这样就形成了包含节点和连边的社交网络图。

在本文的实验中，因为节点数太多，所形成的网络拓扑图较稀疏，所以在对语料进行预处理时，剔除掉节点度小于 3 的节点。在实际的意义中，就是将写著论文数少于 3 的作者删除，这样的作者不算是多产的作者，所以以后继续合著的几率也较小。

将数据集分为10份，9份训练集和1份测试集，并进行交叉验证，得到平均的结果。

通过以上对语料的处理后得到实验的数据集如下表所示。

表 4.1 语料处理后的数据集

Tab.4.1 The dataset after process to the corpus

	Train		Test
	Authors	Edges	Edges
CA-hepth	7734	27662	13831
CA-astro	17290	257768	12884
CA-condmat	20221	112950	56475
CA-grqc	3809	15485	7742
CA-hepph	10495	151634	75816

4.3.2 实验流程

文中选取了 5 种不同的论文合著网络，即 CA-hepth，CA-astro，CA-condmat，CA-qrqc 和 CA-hepph.不同的数据集虽然都是论文合著数据集，但是涉及到不同领域的网络图，让数据集的选择既具有相似可比性，又有其自身的特点，减少了实验的偶然性。

在基于 AdaBoost 提升算法进行链接预测的研究中，将链接预测的问题看作是一个多分类的问题，每一个算法都可以看作是一个分类器。基于 AdaBoost 提升算法最终得到的是强分类器，而强分类器是由弱干弱分类器按比例组合而成，从不同的角度，对链接预测算法进行分类，将各个分类器看作是弱分类器。

在选取弱分类器时，将链接预测算法分为两类：基于邻近度方面的算法和基于中心度方面的算法。而两类不同的算法中选取了若干代表算法，为每一个节点赋予一个权重，通过每一次弱分类器的分类准确率，提高分类错误的节点权重，使其在下一个分类器中更容易被识别出来，改变了数据集的分布情况，并根据分类准确率情况得到该分类器在强分类器中所占的比重。

不仅仅得到了所有弱分类器按比例组合成的强分类器，也根据不同弱分类器的结果，找到在两类不同的分类器中，结果最好的分类器，即有两个最好的分类器分别代表不同类别的分类器。将这两个分类器进行组合，得到另一个强分类器。

文中得到的所有结果在评价体系中进行判别，在图形和表格中对各个结果进行比较，并分析实验的最终结果。

4.3.3 实验评价指标

在对链接预测的评价指标中，一般的评价指标有准确率，召回率，F值，ROC曲线和AUC值等几个方面，不同的评价指标侧重点也有所区别。

准确率也叫精度Precise，文中简称P，在不同的应用中，也有不同的评价标准。在测量等计量方法中，准确率反映了测量值和真实值的误差。在检索等应用领域，主要是检索到的文档数和被检索相关文档数的比值，本文主要是指链接预测对的节点对数和所有链接预测节点对数的比值。准确率反映了分类器在预测的所有链接预测中正确的链接。具体公式如下所示：

$$P = \frac{\text{链接预测正确的节点对数}}{\text{所有链接预测的节点对数}}$$

召回率也叫查全率recall，文中简称R，在检索等应用领域，召回率是指检索出的文档数和文档集中所有文档数的比值。文中的召回率是链接预测正确的节点对和测试集中节点对的比值，召回率反映了分类器的灵敏度和完整性。

$$R = \frac{\text{链接预测正确的节点对数}}{\text{测试集中所有节点对数}}$$

而一般准确率和召回率相互制约，而F值可以对两者有一个平衡的作用，具体公式如下所示：

$$F = \frac{2 * P * R}{P + R}$$

只有当P和R的值同时都很高时，得到的F值才会很高。

ROC曲线反映了分类器的敏感程度，对分类正确和分类错误的惩罚不一样，曲线由FPR和TPR两个量之间的比值变化得到，其中FPR(false positive rate)是指将错误的分类数看作是正确的分类数的比值，TPR(true positive rate)是指将正确的分类数的比例。ROC曲线中主要计算曲线下的面积，曲线的面积和分类器的敏感程度成正相关。

为了对ROC曲线下的面积值有更直观的数值分析，用AUC的值表示ROC面积值。AUC的值介于0.5到1.0之间，通常AUC的指越大，反映分类器越灵敏。

不同的衡量指标从不同的角度对分类器进行了刻画，全方面的了解分类的效果，更客观的评价分类结果。

4.3.4 实验结果及分析

由实验流程，在弱干个弱分类器中进行实验，同时将弱分类器线性组合成强分类器，在两类最好的链接预测算法中，得到效果最好的链接预测算法，并将这两个弱分类器组合成强分类器 1，在评价体系准确率、召回率、F 值、ROC 曲线和 AUC 值中，结果如下所示。

表 4.2 准确率

Tab.4.2 The precision rate

	CN	Jaccard's	AA	EXP	NEU	PA	PR	强分类器 1	强分类器
CA-hepth	0.2258	0.1069	0.1551	0.0028	0.0036	0.0014	0.0001	0.2312	**0.3268**
CA-astro	0.0893	0.0069	0.0548	0.0011	0.0024	0.0013	0.0013	0.0897	**0.1002**
CA-condmat	0.2233	0.1152	0.1414	0.0032	0.0044	0.0017	0.0017	0.2303	**0.2788**
CA-grqc	0.2249	0.1211	0.1504	0.0043	0.0050	0.0036	0.0001	0.2369	**0.3541**
CA-hepph	0.0777	0.0409	0.0497	0.0011	0.0019	0.0022	0.0001	0.0805	**0.1067**

表 4.3 召回率

Tab.4.3 The recall rate

	CN	Jaccard's	AA	EXP	NEU	PA	PR	强分类器 1	强分类器
CA-hepth	0.4515	0.2138	0.3102	0.0031	0.0047	0.0029	0.0001	0.4566	**0.4899**
CA-astro	0.1786	0.0914	0.1097	0.0027	0.0033	0.0026	0.0026	0.1896	**0.2063**
CA-condmat	0.4465	0.2304	0.2827	0.0030	0.0044	0.0034	0.0034	0.4598	**0.5639**
CA-grqc	0.4499	0.2422	0.3008	0.0033	0.0048	0.0071	0.0017	0.4591	**0.5203**
CA-hepph	0.1554	0.0818	0.0995	0.0014	0.0021	0.0045	0.0015	0.1609	**0.1792**

表 4.4 F 值

Tab.4.4 The value of F

	CN	Jaccard's	AA	EXP	NEU	PA	PR	强分类器 1	强分类器
CA-hepth	0.3010	0.1425	0.2068	0.0029	0.0041	0.0019	0.0001	0.3070	**0.3921**
CA-astro	0.1191	0.0128	0.0731	0.0016	0.0028	0.0017	0.0017	0.1218	**0.1349**
CA-condmat	0.2978	0.1536	0.1885	0.0031	0.0044	0.0023	0.0023	0.3069	**0.3731**
CA-grqc	0.2999	0.1615	0.2005	0.0037	0.0049	0.0048	0.0002	0.3125	**0.4214**
CA-hepph	0.1036	0.0545	0.0663	0.0012	0.020	0.0030	0.0002	0.1073	**0.1338**

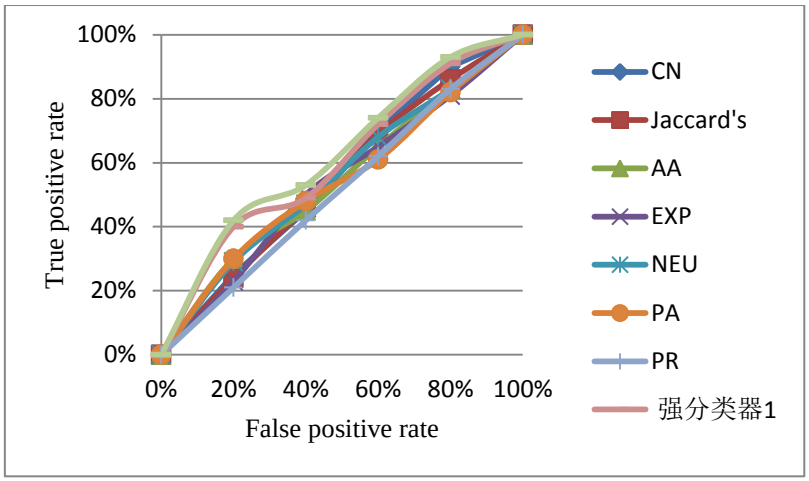

Fig.4.1 The ROC value in the date set of CA-hepth

图 4.1 数据集 CA-hepth 中的 ROC 曲线

Tab.4.5 The value of AUC in the data set of CA-hepth

表 4.5 数据集 CA-hepth 中的 AUC 的值

	CN	Jaccard's	AA	EXP	NEU	PA	PR	强分类器 1	强分类器
CA-hepth	0.57	0.56	0.58	0.56	0.58	0.51	0.51	0.61	**0.67**
CA-astro	0.63	0.55	0.59	0.53	0.56	0.52	0.54	0.65	**0.68**
CA-condmat	0.55	0.53	0.61	0.51	0.55	0.53	0.53	0.63	**0.65**
CA-grqc	0.58	0.52	0.53	0.52	0.51	0.51	0.51	0.64	**0.65**
CA-hepph	0.51	0.51	0.51	0.55	0.52	0.53	0.52	0.60	**0.63**

以上表格和数据直观的显示了实验的结果，通过不同方面的比较，可以得知各个算法的优劣，不同的数据集也说明了实验并非偶然形成。

上表 1，表 2 和表 3 分别表示在不同数据集中的准确率、召回率和 F 值。可以看出将所有的弱分类器组合成强分类器后，无论是准确率、召回率还是 F 值均比其它算法大。

两类分类效果最好的弱分类器分别是 CN 和 PA，将这两类弱分类器组合成强分类器 1，得到的结果比其它弱分类器好，但比强分类器略差。

图 1 和表 4 分别表示在数据集 CA-hepth 中的 ROC 曲线和 AUC 的值，其中效果最好的是强分类器，其次是强分类器 1。

根据以上结果得知，在不同的数据集中，将所有弱分类器组合成强分类器比单独的弱分类器，无论是准确率、召回率还是 AUC 值等都更好。这也说明了基于 AdaBoost 算法的正确可靠性。

4.4 本章总结

本章节主要研究了社交网络中节点之间的关系，即节点链接预测。不仅对以往链接预测的算法进行了总结，也提出了新的预测的方法，并对最终的结果在不同的评价指标中进行了评价。

将链接预测算法分为两类：基于邻近度方面的链接预测算法和基于中心度方面的链接预测算法。而这两个方面中，前者主要侧重于从邻居节点中找到节点的属性特征，找到有链接的节点对之间的节点所具有的特征，进行节点预测和推荐；而后者主要是从社交网络的网络拓扑结构进行研究，从某一个节点出发，按照随机游走的方式，有更大的概率游走到某节点，那么该节点就是预测或推荐的节点。

为了综合利用不同类型的链接预测算法，根据统计学习的方法，提出了基于 AdaBoost 提升算法的链接预测算法，将多种已有的链接预测的算法看作是弱分类器，多个弱分类器按比例的加权组合，给节点赋予权重，通过改变数据集的分布，得到强分类器。文中强分类器表示基于所有的弱分类器得到，而强分类器 1 是基于两类中分类效果最好的弱分类器，即 CN 和 PA 组合得到，由实验结果得知强分类器的效果最好，强分类器 1 效果次之，这也说明了基于 AdaBoost 算法的链接预测的正确性和有效性。这和社会经验中，“三个臭皮匠赛过诸葛亮”相似，多个众人得出的结论可能会比三两个专家要好，而 AdaBoost 算法也正是“集体智慧”的体现。

结　论

在线社交网络的快速发展，也催生了对社交网络的研究和应用。不同于以往对社交网络的研究，大多侧重于社交网络结构、节点或边的分布等的研究，本文抽取社交网络中节点和边，分析节点和边的研究情况。

在社交网络节点的研究中，选取微博社交网络，估量微博中用户节点的影响力情况。通过结合微博中用户评论、转发等行为，给每一个用户赋予一个权重，在微博社交网络中基于随机游走算法，得到每一个用户的影响力值。在真实数据和新浪微博数据中进行了实验，验证了该方法更能客观真实的反映用户的实际影响力。

在针对社交网络中节点关系的研究中，文中选取的是节点的链接预测。总结以往对链接预测的相关研究，并将弱干个链接预测算法看作是弱分类器，通过改变数据集的分布情况，将弱分类器按比例线性组合成强分类器。在多个合著网络和不同的评价指标中，说明了基于 AdaBoost 的链接预测算法可以更好的预测节点的链接情况。

在未来可以从更细致的方面进行研究。比如在对社交网络中节点影响力的研究中，可以根据不同的话题，或加入时间因素等，分析随着话题的不同，随着时间的推移，节点影响力的变化情况。

在对社交网络中节点关系的研究中，可以将边赋予不同的权重，预测节点链接或者不链接的概率等。

参 考 文 献

[1] Kossinets G, Watts DJ.Empirical analysis of an evolving social network[J]. Science. 2006,311(5757):88-90.

[2] David Liben-Nowell, Jon Kleinberg. The link-prediction problem for social networks [J].Journal of the American Society for Information Science and Technology.2007.58(7):1019-1081.

[3] 东昱晓,柯庆,吴斌.基于节点相似性的链接预测[J].计算机科学. 2011.38(7):162-164.

[4] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto etc. Measuring User Influence in Twitter: The Million Follower Fallacy[C].Association for the Advancement of Artificial Intelligence(AAAI 2010).Atlanta, Georgia, USA. 10-17.

[5] 石磊，张聪，卫琳.引入活跃指数的微博用户排名机制[J].小型微型计算机系统，2012(1):110-114.

[6] Danah Boyd，Scott Golder，Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter[C].Institute of Electrical and Electronic Engineers(IEEE 2010).Harbin.4-10.

[7] 张华平，孙梦姝，张瑞琪等.微博博主的特征与行为大数据挖掘[J].中国计算机学会通讯，2014(6):36-43.

[8] Y. Yamaguchi, “TURank: Twitter User Ranking Based on User-Tweet Graph Analysis”, Web Information Systems Engineering, Vol. 6488, 2010, pp. 240-253.

[9] Weng J S, Lim E P, Jiang J, et al. TwitterRank: Finding Topic-sensitive influential Twitterers[C]. In: Proceed-ings of the 3rd ACM International Conference on Web Search and Data Mining(WSDM 2010). New York, USA. ACM 2010:261-270.

[10] 王琛,陈庶樵.一种改进的微博用户影响力评价算法[J].信息工程大学学报,2013(6):380-384.

[11] 徐建.基于复杂网络的节点影响力评估模型[J].软件导刊.2014,13(3):42-45.

[12] Shaozhi Ye, S. Felix Wu.” Measuring message propagation and social influence on Twitter.com”,SocInfo 2010, LNCS 6430, pp, 216-231.

[13] Daniel M. Romero, Wojciech Galuba, Sitaram Asur etc. “Influence and passivity in social media”, Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science, Vol. 6913, 2011, pp. 18-33.

[14] 谢世娜，李川.基于用户复杂联系的最大影响力节点发现算法[J].现代计算机：专业版.2015,03,002.

[15] 郭静，曹亚男，周川等.基于线性阈值模型的影响力传播权重学习[J].电子与信息学报.2014,36(8):1804-1809.

[16] 卢体广，刘新，刘任任.微博数据通用抓取算法[J].计算机工程.2014,05,003.

[17] 黎明,文海英,杨杰等.基于行为权值的微博用户影响力度量算法[J].计算机工程与应用.2014,(17):130-133.

[18] Romero D M,Galuba W,Asur S,et al. Influence and passivity in social media[C].The 20th International Conference Companion on World Wide Web(WWW’11).hyderabad. India,March 2011:13-114.

[19] Ding Z Y,Jia Y,Zhou B.et al. Mining topical influencers based on the multi-relational network in micro-blogging sites[J].Communications China.2013,10(1):93-104.

[20] 刘继，李磊.面向舆论主题的微博用户网络影响力挖掘分析[J].情报杂志.2014,(10):21-24.

[21] 吴渝，马璐璐，林茂等.基于用户影响力的意见领袖发现算法[J].小型微型计算机系统.2015,(3):561-565.

[22] David Liben-Nowell, Jon Kleinberg. The link prediction problem for social networks[C]. Proceedings of the 12nd ACM International Conference on Information and Knowledge Management(CIKM’03). New Orleans, USA. 2003, 556-559.

[23] Liyan Dong, Yongli Li, Han Yin, etc. The algorithm of link prediction on social network. Mathematical Problems in Engineering[J], vol. 2013, Article ID 125123, 7 pages, doi:10.1155/2013/123123.

[24] 吕琳媛，周涛.链路预测[M].高等教育出版社.2013.

[25] L.Lv, T.Zhou. Link Prediction in Complex Networks: A Survey[J]. Physics. 2010,1-44.

[26] 刘宏鲲，吕琳媛，周涛.利用链接预测推断网络演化机制[J].中国科学：物理学 力学 天文学.2011,41(7):816-823.

[27] 吕琳媛.复杂网络链接预测[J].电子科技大学学报.2010,39(5):651-661.

[28] Tao Zhou, Linyuan Lv, Yi-Cheng Zhang. Predicting missing links via local social networks[J]. The European Physical Journal B. 2009,71:623-630.

[29] Linyuan lv, Yi-Cheng Zhang, etc. Leaders in Social Networks, the Delicious Case. In: PloS ONE, 2011, vol. 6, no.6, p.e21202.

[30] Dong Li, Zhiming Xu, Sheng Li, etc. Link prediction in social networks based on hypergraph[C]. Proceedings of the 22nd International World Wide Web Conference Committee(IW3C2). Rio de Janeiro, Brazil.2013, 41-42.

[31] Zhijun Yin, Manish Gupta, Tim Weninger, etc. A unified framework for link recommendation using random walks[C]. Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining(ASONAM 2010). Odense, Denmark. 2010, 152-159.

[32] Jerome Kunegis, Julia Preusse, Felix Schwagerit. What is the added value of negative links in online social networks?[C]. Proceedings of the 22nd International World Wide Web Conference Committee(IW3C2). Rio de Janeiro, Brazil.2013, 727-736.

[33] 张珊靓，周宴.基于随机游走的时间加权社会网络链接预测算法[J].计算机应用与软件.2014,31(7):28-30.

[34] 郭景峰，代军丽，马鑫等.针对通信社会网络的时间序列链接预测算法[J].计算机科学与探索.2010,4(6):552-559.

[35] Jiawei Zhang,Xiangnan Kong,Philip S.Yu.Transferring heterogeneous links across location-based social networks[C]. 2014:303-312.

[36] Zhaochen Guo, Denilson Barbosa. Robust Entity Linking via Random Walks[C]. Proceedings of the Conference on Information and Knowledge Management(CIKM). Shanghai, China.2014, 499-508.
[37] 潘果，徐雨明.LBSN中位置信息与网络拓扑相融合的好友预测[J]. 计算机科学. 2014, 41(9): 115-118.
[38] 李旺龙，李川.基于用户质量的关注关系预测[J].现代计算机：专业版.2015,03,001.
[39] Hasan M A, Chaoji V, Salem S, et al. Link prediction using supervised learning[J]. Proc of Sdm Workshop on Link Analysis Counterterrorism & Security. 2006.
[40] Gong N Z, O Joint. Link Prediction and Attribute Inference using a Social-Attribute Network[J]. ACM Transactions on Intelligent Systems & Technology. 2014,5(2):529-544.
[41] 杨珺，杨炳儒，唐志刚.基于半监督学习的链接预测算法的研究[J].计算机应用研究. 2010, 27(8):2848-2852.
[42] Menon A K, Elkan C. Link prediction via Matrix Factorization[J].Lecture Notes in Computer Science. 2011,6912(1):437-452.
[43] 张玉芳,孔润,熊忠阳等.Markov逻辑网在链接预测中的应用[J].计算机应用研究. 2011, 28(6): 2154-2157.
[44] John S.Breese,David Heckerman,Carl Kadle. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[R]. Technical Report. 1998,1-20.
[45] 朱索格，胡访宇.基于海量数据的链接预测方法研究[J].电子技术.2014,03,008.
[46] 仇丽青，陈卓艳.社会网络链接预测算法研究[J].软件导刊.2014,(10):61-62.
[47] 卢文羊，徐佳一，杨育彬.基于LDA主题模型的社会网络链接预测[J].山东大学学报：工学版. 2014, 44(6):26-31.
[48] Cainelli G, Maggioni MA, Enka Uberti, et al. The strength of strong ties: How co-authorship affect productivity of academic economists?[J]. Scientometrics. 2015, 102(1): 673-699.
[49] Apte E, Bibelnieks, R. Natarajan, E.Pednault, et al. Segmentation-based modeling for advanced targeted marketing[C]. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD). 2001, 408-413.
[50] Lugosi N C A G. Potential-based Algorithms in On-line Prediction and Game Theory[J]. Machine Learning. 2003, 51(3):239-261.
[51] 吴祖峰, 梁棋, 刘峤等.基于AdaBoost的链路预测优化算法[J]. 通信学报, 2014, 35(3): 116-123.

攻读硕士学位期间发表学术论文情况

1 吴慧，张绍武，林鸿飞.基于微博社交网络的用户影响力评价[J].中文信息学报.2015.（已录用，本硕士学位论文第 3 章）

致　　谢

时光匆匆，聚散容易。转眼间，我的研究生阶段也即将结束，19 年的校园求学生活也画上了句号。借此毕业论文完成之际，向所有在我成长道路中帮助、鼓励我的人献上最诚挚的敬意和感谢。

首先，感谢我的导师张绍武。谢谢您将我带入信息检索实验室，开启了一段美好的学术之旅，让我接受了全新的思想观念，扩宽了我对学术领域的思考空间。您真诚朴实、乐观积极的生活态度，也让我受益匪浅。同时，也感谢林鸿飞教授。在 IR 实验室，无论是新生徒步走，还是新年晚会上各种精彩的节目，或者实验室的羽毛球比赛，亦或女生节的鲜花，都让我刻骨铭心。您在学术领域里，精益求精、视野广阔、严谨认真的学术作风让我受益无穷。在生活中，您的细致和人文关怀，如沐春风，倍感温馨。也希望多年后能够像您一样，睿智的工作和生活。

其次，感谢实验室里的同学，在一起组会时的激烈讨论，各种活动时留下的欢笑，非常感谢你们的陪伴。特别感谢 B907 的师兄弟姐妹们，大家在每天的朝夕相处中，讨论学术，也商讨生活的点滴，让我非常感动。刚进入实验室的时候，就是一张白纸，不会编程，不会写代码，对信息检索没有任何的概念，非常迷茫，感谢任克江、李学妮、魏现辉等师兄师姐的帮助和鼓励，让我逐渐的进步。感谢我的好朋友赵虹杰，潘文慧，非常感谢她们对我的陪伴和鼓励。大家在一起分享快乐、分担忧愁，希望我们都有一个美好的未来。

感谢学校和实验室，给予我成长的环境，为我提供了学习和进步的平台，让我学会感恩，感激今天所得到的许多，让我倍感珍惜，也让我对以后的工作和生活充满热情和信心。

再次，感谢我的父母，谢谢你们无私的付出，让我在一个温馨美好的家庭中成长，能够有条件心无旁骛的学习，我无以为报，只有更加努力的工作和学习，让你们放心。你们身体健康是我最大的安慰。同时感谢我的哥哥，谢谢他在生活中对我的关怀和照顾，此外，谢谢小侄女给我们带来的欢乐，希望她能健康快乐的成长。

最后，衷心感谢评审本论文的各位老师。

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：____年____月____日

导 师 签 名：_____ 日期：____年____月____日