时变网络的链接预测研究

重庆大学硕士学位论文

（学术学位）

学生姓名：武南南

指导教师：邢永康　副教授

专　　业：计算机软件与理论

学科门类：工　学

重庆大学计算机学院

二〇一二年四月

Research on Link Prediction Based on Time-Varying Networks

A Thesis Submitted to Chongqing University
in Partial Fulfillment of the Requirement for the
Master's Degree of Engineering

By

Wu Nannan

Supervised by Ass. Prof. Xing Yongkang

Specialty: Computer Software and Theory

College of Computer Science of
Chongqing University, Chongqing, China
April, 2012

摘　要

链接预测是链接挖掘的一个分支，主要是基于对象的属性和其他观测到的链接，预测两个对象之间是否存在链接。链接预测算法，可以用于发现丢失的信息、预测未来将要发生的事件、评估网络的演化机制等。链接预测研究，对于许多当前社交网络中比较流行的应用有着重要的影响。例如：链接预测在预测社交网络的丢失信息上，扮演着重要角色。人工智能和数据挖掘领域的研究者认为，一个像公司那样的庞大组织，能从分析员工的非正式社交网络数据中获益。高效的链接预测算法，可以被用于分析像社交网络类的时变网络，并且获得具有一定可信度的结论。

在具有多样结构、混杂和无规律的复杂网络中，传统的数据挖掘方法是无法应付的。如果，仅仅将基于独立同分布假设的传统的数据挖掘方法应用于这些数据集，挖掘出来的结论是不合适的。因此，在面临这类复杂的网络时，我们必须细心关注和利用那些潜在的链接关系以及对象之间动态变化的关系，挖掘的结果才是合适的。实际上，对象之间的链接关系也是一种知识，在进行数据挖掘时，我们应该充分利用这些知识。所以，在本文中，我提出了时变网络的动态演化模型来精确量化对象之间的关系，改进了传统的链接预测算法以适应于动态演化模型，并且结合马尔科夫逻辑网提出了一个新的链接预测算法。根据在 Enron 数据集上的实验结果，本文改进的链接预测算法和提出的新的链接预测算法均优于传统的链接预测算法。

在本文中，我的主要研究内容包括如下三部分。

① 本文提出一种描述社交网络等时变网络的演化过程的动态演化模型。传统的时变网络的静态模型只是简单统计对象之间是否有事件发生，而没有精确描述对象之间的关系随时间的变化过程，但是时变网络的动态演化模型不仅包括了静态模型所表达的信息，还引入了对于对象之间的关系的变化有深刻影响的时间序列。

② 本文改进了一些传统的链接预测算法，以适应于时变网络的动态演化模型。经过改进的链接预测算法，对时变网络的链接预测准确率有明显地提高。

③ 根据马尔科夫逻辑网模型的特性，本文提出了一种新的链接预测算法。由于每个传统的链接预测算法在不同性质的数据集上，链接预测结果相差很大，甚至不同的算法在一个模型中的预测结果是截然相反的，然而马尔科夫逻辑网模型可以很好地兼容不同的链接预测算法，甚至是互斥的算法，

所以结合传统的链接预测算法和马尔科夫逻辑网模型，提出了一个新的链接预测算法。新的链接预测算法在时变网络中的效果明显优于传统的链接预测算法。

关键词：链接预测，马尔科夫逻辑网，时变网络，链接挖掘，数据挖掘

ABSTRACT

Link prediction is a branch of the link mining. According to attributes of the objects and other observed links, link prediction is the problem of predicting the existence of a link between two objects. The algorithms of link prediction can be used to extract the missing information, to predict the incidents occurring in the future and to evaluate the mechanism of the evolution of networks. The research of link prediction plays an important role in the current popular applications of social networks. For example, link prediction plays an irreplaceable role in predicting the missing messages of social networks. Researchers in artificial intelligence and data mining have argued that a large organization, such as a company, can benefit from the interactions within the informal social network among its members. Effective methods for link prediction could be used to analyze social networks to get some credible conclusions.

A key challenge for traditional data mining is tackling the problem of mining richly structured, heterogeneous, non-law networks. Naively applying traditional data mining approaches based on the IID assumption can lead to inappropriate conclusions about the data. So, when it comes to these networks, we must take account of potential correlations due to links and the relationship between objects changing over time, for getting the appropriate mining results. In fact, object linkage is also knowledge that should be exploited when other knowledge is mined from networks. So, Time-Varying Network Model is proposed to quantify the relationship between objects in this paper. The traditional algorithms of link prediction are improved to be suitable for Time-Varying Network Model. And one novel algorithm of link prediction based on Markov Logic Network is proposed in this paper. According to the experiment results on Enron dataset, the improved algorithms of link prediction and the novel algorithm of link prediction are both better than the traditional algorithms of link prediction.

In this paper, my main research topics include the following parts.

① Time-Varying Network Model for social networks is proposed in this paper. The traditional static graph of time varying network is just a simple representation of events occurred between objects, rather than quantifies the process of relations between objects changing over time precisely. But

Time-Varying Network Model not only includes the expression of the static graph of time varying network, but also introduces a time series that have a profound impact on the relationship between objects changing over time.

② The traditional algorithms of link prediction are improved to be suitable for Time-Varying Network Model. The improved algorithms of link prediction for the time-varying network link prediction accuracy have significantly improved.

③ According to the attributes of Markov Logic Network, a novel link prediction algorithm is proposed in this paper. We propose the novel algorithm of link prediction that combines Markov Logic Network with the traditional algorithms of link prediction, for the reason that the results of traditional algorithms of link prediction in the dataset of different nature are very different, and even the results of different algorithms in the same model are diametrically opposed, however, Markov Logic Network can be compatible with different algorithms of link prediction and even the exclusive algorithms. The novel algorithm of link prediction for the time-varying network based on Markov Logic Network and the traditional algorithms of link prediction is much better than the traditional algorithms of link prediction.

Keywords: Link Prediction, Markov Logic Network, Time-varying Network, Link Mining, Data Mining

目　　录

1 绪　论

1.1 问题提出与研究意义

1.1.1 问题提出

链接预测的主要问题就是基于对象的属性和其他观测到的链接，预测两个对象之间是否存在链接[1]。数据之间的“链接”，更宽泛地讲为关系，无处不在。这些链接通常表示为数据实例之间的一些模式，例如，对象的重要性等。在目前许多实验用的数据集上，链接使得对象之间相互关联，并且以链接集合的形式存放。在某些数据集中，有些链接是丢失的，有些链接是在将来发生的，并且链接的发生是一个动态过程，而不是所有的链接都可以被观测到。因此，对于预测实例之间存在的链接引起我们极大兴趣。

传统的链接预测方法都是，假设对象之间的关系是静态的，是不会随时间而变化的。然而，在一些实际应用中，这个假设明显是错误的。例如，在学术论文的引用中，作者与作者之间只有随时间的不断相互引用，才会加强他们之间的关系，如果他们之间不再有引用关系，则他们之间的关系也相应减弱了。目前，一些链接预测方法都是基于统计对象之间是否发生事件的网络，在本文中，该类网络称为静态图，因为该类网络丢失了网络中事件发生的时间特性。例如，在电子邮件网络中，发送和接收邮件的时间序列决定了对象与对象之间的关系程度。在一些社交网络中，链接是随着时间在变化的，结合已观测到的链接，我们的目标是去预测一个链接是否会在将来发生。

同时，传统的链接预测方法在同一个数据集上的预测结果相差很大，甚至相反，但是，经实验证明，传统的链接预测方法在一适合的数据集上，预测结果极好。然而，基于传统模型的链接预测方法不能融合各个传统链接预测方法的优点，提高链接预测的准确率。

在传统的数据挖掘方法中，我们只是学习一个只考虑复杂网络中节点的属性，而忽略了节点之间链接的关系。传统的数据挖掘方法，无法处理具有丰富、多样结构、混杂，无规律和动态变化的数据集，这种数据集通常表示成一个网络图，一个随时间变化的网络。如果只是应用基于独立同分布假设的传统数据挖掘方法在这些数据集上，挖掘出来的结论是不适合的。在这样的数据集进行挖掘时，我们必须细心关注那些潜在的链接和对象之间发生事件的序列，得到的结果才是合适的。实际上，对象之间的链接关系也是一种知识，在进行数据挖掘时，我们应该利用这些知识。

然而，由于许多数据集的链接都非常稀疏，导致很难通过链接预测方法

得到一个好的预测结果。许多研究者认为，建立链接预测统计模型的一个问题是链接的先验概率非常小[2,3,4]。如何，提高链接预测质量成为了许多研究者的主要目标。在本文中，一个提高链接预测质量的方法是综合预测。另外一个问题是，传统的链接预测方法都是基于时变网络静态图模型。然而，在时变网络静态图中，忽视了一个对链接预测影响非常重要的时间因素。

1.1.2 研究意义

链接预测研究有极其广泛的研究意义。链接预测研究对于许多当前社交网络中比较流行的应用有着重要的影响[5]。例如：链接预测在预测社交网络的丢失信息上，扮演了重要角色。人工智能和数据挖掘领域的研究者认为，一个像公司那样的庞大组织，能从分析员工的非正式社交网络数据中获益；这些从组织本身挖掘出的信息可以直接辅助管理层决策[6,7]。高效的链接预测算法能被用于分析像社交网络这样的时变网络，得出具有一定可信度的结论，例如：预测出该组织中还没有发生，但即将发生与其它组织的合作关系。

在另外一个领域中，受到分析恐怖组织社交网络问题的刺激，社交网络分析中的安全领域方向的研究越来越受到重视。在这样的背景中，某人即使没有和极端分子直接地交往，但是根据他的社交关系和发生事件的序列，链接预测也允许我们推断出这个人参与了他们的活动[8]。链接预测可以，在作者的合著关系网中预测两个作者之间新的合著的可能性[5]，改进搜索引擎的超文本分析方法，挖掘犯罪分子组成的社会网络中隐藏的重要关系等。

1.2 国内外研究现状

链接预测的研究交叉于链接分析、超文本挖掘、web 挖掘和关系学习等领域，是最近几年出现的一个新的研究领域。目前对于链接预测的国内外研究主要可以分为两类。

第一类是基于网络结构的链接预测方法。在这种方法中，利用对象之间的链接关系形成的网络结构信息。根据预测问题，选定不同的亲近度计算公式，通过亲近度来预测对象之间出现链接的可能性。链接预测算法常用的亲近度如下。一、基于两个对象之间距离集合[5]。两个对象之间的最短距离越短，则发生链接的可能性越大。这类亲近度计算公式简单，没有充分利用链接图的信息，链接预测的准确率不高。其中，具有代表性的链接预测算法：SimRank、Hitting time 和 Katz 等方法 [5,9,10]。二、基于两个对象之间的共同邻居[5]。这类亲近度计算公式主要有 Common neighbors、Jaccard's coefficient、Adamic/Adar 等[11,12]，在存在大量链接信息的图中表现卓越。M. Craven 等建立了一个关于 World Wide Web 的本体库，包含各种超链接关系，此模型可以，

应用到知识推理和链接预测等方向[13]。O'Madadhain 等建立了一个基于时间序列的对象排序模型，在此模型上可以动态地进行对象的重要性排序[14]。

第二类是基于概率关系模型(Probabilistic relational models: PRMs)的链接预测方法。这种方法一般根据机器学习算法，建立对象的属性和链接之间的模式映射关系，从而进行链接预测。Popescul 等利用关系特征，构造结构化逻辑，通过该模型去预测链接的存在性[16]。这个关系特征通过数据库查询语言定义[17]。O'Madadhain [15]等构造了基于属性和结构属性的局部条件概率模型(Local conditional probability models)，一个链接是否发生取决于条件概率的大小。这类方法一般是，建立一个关于整个链接网络的统一的概率模型。该模型是一种统计关系模型，在该模型中集合了多种策略用于链接预测。这些概率模型通常都是基于马尔科夫随机场(Markov random fields)[18]。Lise Getoor[19]等通过概率来量化属性和链接之间的相互作用，并且提出引用的不确定性和存在的不确定性两种机制来表示在链接图上的概率分布，建立了一个概率关系模型。这类方法目前研究广泛，存在的主要问题：在模型的建立中许多链接的先验概率难以获得，预测时计算量较大。

1.3 本课题研究的主要内容

本课题研究的主要目的：发现一种拥有较高准确率、执行时间较短、可以适应不同种类时变网络(如：社交网络)的链接预测算法，以充分利用时变网络中的链接信息、节点属性和时间因素等特性得到良好的预测结果。在本文中，提出一个对社交网络等时变网络随时间演化的动态演化模型，并且根据该模型改造一些对不同性质的数据集各有不同优缺点的经典链接预测算法，研究利用马尔科夫逻辑网模型，结合经过改进的链接预测算法，研究出一种新的链接预测算法。

本文主要对链接预测进行研究，围绕时变网络、马尔科夫逻辑网展开。首先，对链接预测的背景和研究现状进行了论述，分析了链接预测在数据挖掘领域、现实生活、甚至生物学科等一些交叉领域中的研究意义，并且对于涉及到的相关理论技术和该领域的发展现状进行了详细探讨，对于了解本文的研究课题打下了一个全面的基础。其次，在本文中，给出了一个对于社交网络等时变网络随时间演化的动态演化模型和在该模型上改造一些对于数据集的不同性质有着不同优缺点的经典链接预测算法，以及将马尔科夫逻辑网模型应用到链接预测。本文中的动态演化模型拟合了现实生活中的社交网络中的关系变化过程。在马尔科夫逻辑网模型和动态演化模型的基础上，本文提出了改进的链接预测算法，并且融合了时间特性解决了传统链接预测算法

没有考虑时间序列的问题。最后，在本文中主要是对实验进行客观的设计，采用 Enron 数据集，以及将新的链接预测算法与经典链接预测算法进行了比较。

① 由于经典链接预测算法都是基于时变网络的静态模型，本文提出一种对社交网络等时变网络随时间演化的动态演化模型。传统的时变网络的静态模型没有精确表达时变网络所有信息，而时变网络的动态演化模型不仅包括了静态模型表达的信息，还引入了对建立合适的时变网络模型重要因素时间序列。时变网络的动态演化模型主要是改变了在静态模型中，描述对象之间关系的方法，设计了基于时间变化的方法来描述对象之间的关系。

② 由于描述时变网络的模型进行了改进，本文对一些经典链接预测算法进行了改进，适合于时变网络的动态演化模型。经过改进的链接预测算法对时变网络的预测准确率有明显地提高。

③ 对不同性质的数据集各有不同优缺点的经典链接预测算法，融入马尔科夫逻辑网模型，并且基于该模型引入改进的链接预测算法，提出新的链接预测算法。在经典的链接预测算法中，每个算法给出的亲近度相差悬殊，甚至不同的算法在一个模型中是互斥的，然而马尔科夫逻辑网模型可以很好地兼容不同链接预测算法，甚至是互斥的算法，因为在马尔科夫逻辑网模型中，它们只是概率的表示形式。

1.4 论文章节安排

本文主要分为六个章节，各章的具体内容如下：

第一章是绪论部分，对时变网络的链接预测研究的问题和意义进行了详细阐述，并分析了链接预测的发展过程和国内外的研究现状。在此基础上，提出了本课题的研究目的以及对主要研究内容进行了概括介绍，并对本课题中的主要研究内容进行了阐述。

第二章是主要介绍了本文研究所涉及到的一些相关理论和技术。详细介绍了链接预测领域的国内外研究现状，并且引用实验结论对一些链接预测算法进行了详细的分析。本文链接预测研究涉及到的马尔科夫逻辑网，也给出了详细的介绍。

第三章是是主要介绍时变网络的概念，时变网络静态模型和时变网络动态模型的概念。最后，对链接预测研究领域中的链接预测方法，进行了全面的分析。

第四章首先，分析了时变网络静态模型和动态模型的描述能力及优缺点，定义了时变网络的动态演化模型，给出了动态演化模型的算法和实例分析。

其次，我们改进了传统的链接预测方法，以适应于时变网络的动态演化模型。最后，我们提出了一个新的基于马尔科夫逻辑网的链接预测算法，该算法融合了传统链接预测算法的优点摒弃了各自的不足。

第五章是实验与实验分析。在这一章中，设计了实验方法，选择采用了度量链接预测算法优劣的标准度量方法以及对数据集进行了预处理。然后，我对于改进的链接预测算法和提出的新的链接预测算法分别进行了实验对比，在最后对实验结果进行了客观、科学的分析与论述。

第六章是总结与展望。在这一章中，我总结了我的主要研究工作，提出了我完成的主要研究内容以及研究中存在的问题与不足，并且对时变网络的动态演化模型和基于马尔科夫逻辑网的链接预测前景进行了分析与展望。

2 相关理论介绍

链接预测研究主要涉及了链接预测方法的研究。我们对目前具有代表意义的链接预测方法给出了详细的介绍和分析。由于在本文的链接预测研究中，我们提出了基于马尔科夫逻辑网的新的链接预测方法，所以对马尔科夫逻辑网的研究也给出了详细的介绍。

2.1 链接预测

2.1.1 链接预测简介

链接预测的主要问题就是基于对象的属性和其他观测到的链接，预测两个实例之间是否存在链接[1]。例如：在演员的社交网络中，预测演员之间的联系，如：友情关系；预测演员之间是否发生事件，如：发送电子邮件，电话呼叫和共同参演一部影片。链接预测还包括预测像基于网页链接和内容的顾问关系的语义关系[13]。

“链接”，更宽泛地讲，充斥于数据实例的各个地方。这些链接通常表示为数据实例属性的一些模式，例如，对象的重要性、排序和分类等。现在许多实验用的数据集都以相互关联对象的链接集合形式存放。在某些数据集中，不是所有的链接都可以被观测到。这就引起了学者们对链接预测研究的极大兴趣。实际上，对象之间的链接关系也是一种知识，在进行数据挖掘时，我们应该利用这些知识。

链接预测是链接挖掘的一子领域，而链接挖掘是一个最近兴起的研究领域。链接挖掘主要包括，基于链接的对象排序，基于链接的对象分类，组发掘，对象标识，链接预测，子图发现和图分类等研究领域。链接挖掘与传统的数据挖掘主要区别是，链接挖掘主要利用了对象与对象之间的链接知识，然而传统数据挖掘面临着一个巨大挑战是处理具有丰富多样的结构、混杂和无规律的数据集，这种数据集通常表示成一个复杂的网络图。如果仅仅应用基于独立同分布假设的传统数据挖掘方法在这些数据集上，挖掘出来的结论是不合适的。在这样的数据集进行挖掘时，我们必须细心关注那些潜在的链接，得到的结果才是合适的。

许多社交网络、生物个体、信息系统等，都可以很好的被描述成网络，节点代表了个体或生物元素(蛋白质、基因等)[20]、计算机和网络用户等，链接表示成节点之间的相互作用或属性等关系。因此，复杂网络的研究变成了很多自然学科的一个共同课题。学者们把很多精力投放到了去理解网络的演

化过程，拓扑结构和功能之间的关系，网络的特征上。一个重要的与社交网络分析相关的科学研究分支是信息检索，它的目标是从大量文献中找到满足某一信息的无结构性质的材料[21]。这也可以看作单词和文档之间的关系预测，现在已经被延伸到链接挖掘上的一些研究问题中，并且这在链接预测中是最基本的问题，根据观测到的链接和节点的属性试图估计两个节点之间的一个链接存在的可能性。

2.1.2 链接预测算法

在这一节中，我们概括了目前广泛研究的主要链接预测算法[22]。目前的链接预测方法，主要可以分为两类，一是基于网络结构的链接预测方法，这种方法主要根据网络结构信息，建立基于邻近节点亲近度计算的各种链接预测算法，二是基于概率关系模型的链接预测方法，这种方法主要是抽象整个网络结构为一个概率模型，学习各种参数，然后进行链接预测。

第一类，基于网络结构的链接预测算法，主要包括基于亲近度的算法和最大似然方法两类。

最简单的链接预测算法就是基于亲近度的算法。该类基于亲近度的算法，直接定义了x和y之间的亲近度(在文献中，也称为邻近度、相似度等)，并且给予(x,y)赋值S_{xy}。对所有未观测到的链接根据它们的亲近度进行排序，这些越高亲近度的链接，越有可能存在实际的链接。尽管基于亲近度算法的研究简单，但关于该类算法的研究仍是链接预测研究领域的主要研究问题。基于亲近度算法的亲近度计算可以非常简单或非常复杂，尽管该类算法对于某些网络的链接预测效果极差，但是在另一些网络中，该类算法的链接预测效果却非常好。在本文中，亲近度和相似度的概念一样。

基于局部亲近度的链接预测方法介绍。

定义 2.1 $\Gamma(x)$：表示网络中，x的邻居节点集合。

定义 2.2 k_x：表示网络中，x的度。

- Common neighbor(CN)[23]：该方法表示x和y之间的共同邻居数目，x和y之间的共同邻居数目越多，x和y之间发生链接的可能性就越大。

$$S_{xy}=\left|\Gamma(x)\cap\Gamma(y)\right| \tag{2.1}$$

在集合论中，$|Q|$是集合Q的基。A是邻接矩阵，很明显的$S_{xy}=(A^2)_{xy}$。如果x和y直接相连，则$A_{xy}=1$，否则$A_{xy}=0$。我们知道$(A^2)_{xy}$表示了x到y路径为 2 的路径数目。Newman[11]用该方法量化了合著网络中的合著关系，并且展示了在两个作者之间共同邻居的数目和这两个作者将来合作的可能性密切相关。

- Salton Index[24]：定义如下。

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\sqrt{k_x\times k_y}} \tag{2.2}$$

● Jaccard Index[25]：定义如下。

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{2.3}$$

● Sorensen Index[26]：这类方法主要用于生态学学科的研究应用，定义如下。

$$S_{xy}=\frac{2\times|\Gamma(x)\cap\Gamma(y)|}{k_x+k_y} \tag{2.4}$$

● Leicht-Holme-Newman Index[27](LHN1)：这个方法给具有较多共同邻居但又不是拥有最多共同邻居，而期望拥有更多共同邻居的节点对赋值较大。

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{k_x+k_y} \tag{2.5}$$

● Preferential Attachment Index(PA)[28]：这个方法应用于自由扩张网络的主要原理是，一个新的链接连到节点x的概率与节点x的度k_x成正比。这个相似的原理可以用于规模没有扩大的网络，而是在每个时间点上老的链接被移除并且新的链接被产生。一个新的链接连接节点x和y的概率与$k_x\times k_y$成正比。受此原理的影响，这类方法可以定义如下。

$$S_{xy}=k_x+k_y \tag{2.6}$$

该类方法的主要功能意义是广泛地用于量化动态变化网络的链接，例如，像渗透、同步和传输等动态网络。该方法的一个主要特点是，亲近度的计算不需要获取邻居节点的信息，所以具有最小的计算复杂度。

● Adamic/Adar Index[29](AA)：这种方法，改进了只是简单地计算共同邻居数目的算法，对于共同邻居少的链接赋予更大的值。该方法定义如下。

$$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k_z} \tag{2.7}$$

● Resource Allocation Index(RA)：该类方法是从复杂网络的资源动态分配中提出来的。假设节点x和y没有链接，节点x可以借助于节点y的共同邻居，将一些资源发送给节点y。在最简单的情况下，我们假设每个传递者拥有一个单元的资源，所以节点x将会平均地将资源分发给它的邻居节点。同理，节点y可以从节点x和y之间获得资源，定义如下。

$$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k_z} \tag{2.8}$$

我们可以很明显地看到这种方法在计算节点对的亲近度时是对称的，即$S_{xy}=S_{yx}$。同时，我们也注意到虽然 Adamic/Adar Index(AA)方法和 Resource

Allocation Index(RA)方法提出的背景不同，但是这两个方法却很相似。确实如此，他们都对具有大量共同邻居节点的节点对赋值很小，尤其是当k_x很小时，两个方法差别很小。只有当k_x很大时，这两个方法才有明显的差别，即RA 方法要比 AA 方法对大量共同邻居的节点对惩罚大。

Liben-Nowell[5]曾在社交网络中，系统地比较过基于亲近度的链接预测方法。而在本文中，将基于亲近度的链接预测方法在性质完全不同的各种网络中进行比较。这些复杂网络包括“蛋白质-蛋白质”交互网络(PPI)、在网络研究领域的科学家组成的合著网络(NS)、美国西部国家电网(Grid)、一个美国政府的博客交互网络(PB)、一个因特网上的路由网络(INT)和美国航空运输网络(USAir)。

表 2.1 基于亲近度的链接预测方法的 AUC 值[32]

Table 2.1 Similarity-based algorithms of link prediction measured by the AUC value

Indices	PPI	NS	Grid	PB	INT	USAir
CN	0.889	**0.993**	**0.590**	0.925	**0.559**	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	0.933	**0.590**	0.882	**0.559**	0.901
Sorensen	0.888	0.933	**0.590**	0.881	**0.559**	0.902
LHN1	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	**0.590**	0.922	**0.559**	0.925
RA	**0.890**	0.933	**0.590**	**0.931**	**0.559**	**0.995**

根据在这些网络中的实验结果，可知链接预测算法 RA 表现的最优，AA 和 CN 算法的表现次优。PA 算法是这些算法中表现最差的，但我们的主要兴趣在于该算法需要很少的信息量，该算法在数据集 INT 和 Grid 上的表现甚至还不如纯粹的猜测。

基于全局亲近度的链接预测方法介绍。

● Katz Index[10]：这类方法主要是基于网络路径的集合，直接对所有路径进行求和，但是随着路径地增长相应的权重是以指数下降的，所以路径越短，相应的权重就越大。该方法的数学表达式如下。

$$S_{xy}=\sum_{l=1}^{\infty}\beta^{l}\bullet\left|paths_{xy}^{<l>}\right|=\beta A_{xy}+\beta^{2}\left(A^{2}\right)_{xy}+\beta^{3}\left(A^{3}\right)_{xy}+\cdots \tag{2.9}$$

其中$paths_{xy}^{\langle l \rangle}$是节点$x$到$y$所有路径为$l$的路径集合，$\beta$是控制路径权重的一个自由参数(指数下降因素)。我们可以很显然地看到一个很小的β参数使得该算法近似于 CN 算法，因为长路径对链接的权重贡献很小。该方法还可以近似地表达为如下数学式。

$$S = (I - \beta A)^{-1} - I \tag{2.10}$$

我们必须注意β的取值小于矩阵A中的最大值以保证收敛到公式(2.9)。

● Random Walk with Restart(RWR)：这类方法是 PageRank 方法[30]的一个直接应用。随机漫步者反复的从节点x以概率c到一个邻居节点，以概率$1-c$到自己。当随机漫步者走到一个稳定状态时，q_{xy}表示节点x到y的概率。

$$\overrightarrow{q_x} = cP^T \overrightarrow{q_x} + (1-c)\overrightarrow{e_x} \tag{2.11}$$

其中P是转移矩阵，如果节点x和y是有链接的则$P_{xy} = 1/k_x$，否则$P_{xy} = 0$。这个公式还可以如下表述。

$$\overrightarrow{q_x} = (1-c)\left(1-cP^T\right)^{-1}\overrightarrow{e_x} \tag{2.12}$$

根据上面的公式，这个算法可以定义为如下数学式。

$$S_{xy} = q_{xy} + q_{yx} \tag{2.13}$$

其中，q_{xy}表示向量$\overrightarrow{q_x}$的第y个元素。

表 2.2 链接预测方法 Katz 和 RWR 的 AUC 值[32]。

Table 2.2 The Katz and RWR algorithms of link prediction measured by the AUC value.

Indices	PPI	NS	Grid	PB	INT	USAir
Katz	**0.972**	0.988	**0.952**	**0.936**	**0.973**	0.956
RWR	0.968	**0.993**	0.760	0.769	0.959	**0.977**

在实验中，RWR 方法的c参数设置为 0.9。这个实验结果说明了，不同的链接预测方法在不同性质的数据集上的表现也不都一样。在 Grid 数据集上，Katz 方法要比 RWR 方法的预测结果理想很多。

基于最大似然估计的链接预测方法介绍。

这类算法假设网络结构的组成是有组织规则的，其中的详细规则和参数是通过对观测到的结构，使用最大似然方法学习到的。根据这些学到的规则和参数，我们可以计算任一个未观测到的链接发生链接的可能性[31, 32]。

实践证明，许多实际的网络都是以层次管理的，其中的节点可以被分成组，更进一步，组也可以分成更细的组，依次类推。例如：人类的大脑网络结构图，可以分成不同的区域。当仔细研究社交网络和生物网络的内在层次

结构时，我们可以提出一些巧妙的方式去发现丢失链接。Clauset 等[33]提出了一种可以推理网络层次结构的通用技术，并且可以将它应用到链接预测中。

在图 2.1 中，一个网络的层次结构能表示成一个树状图。一个树状图有和网络中节点一样多的 N 个叶子节点，和 N-1 个内置节点。Clauset 等[33]提出了一个简单的模型，在该模型中，每个内置节点r有一个概率p_r，并且每对叶子节点之间的链接发生的概率是$p_{r'}$，而r'是这对叶子节点最小共同祖先。

一个实际网络 G 和一个树状图 D，E_r表示以节点r为最小共同祖先的节点在 G 中组成的边的数目，让L_r和R_r分别表示以节点r构成树中左子树的叶节点数目和右子树的叶节点数目。这个树状图 D 的似然函数如下公式。

$$L(D,\{p_r\})=\prod_r p_r^{E_r}(1-p_r)^{L_r R_r-E_r} \tag{2.14}$$

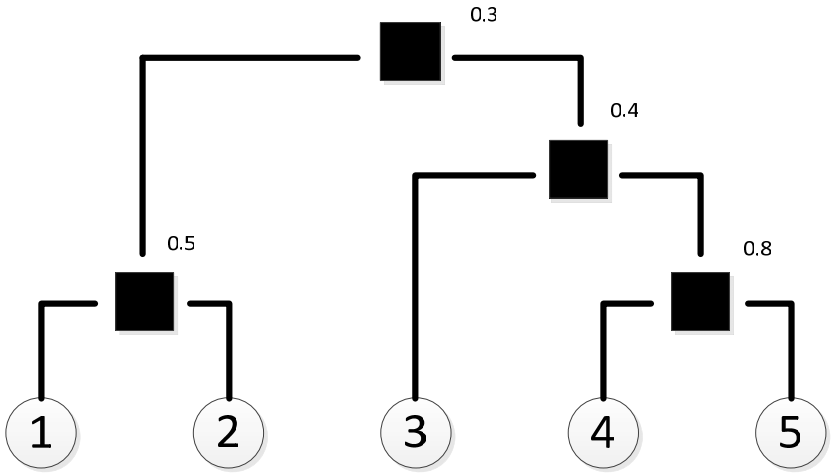

图 2.1 五个节点的网络图的树状图。

Fig 2.1 Illustration of dendrogram of a network with 5 nodes.

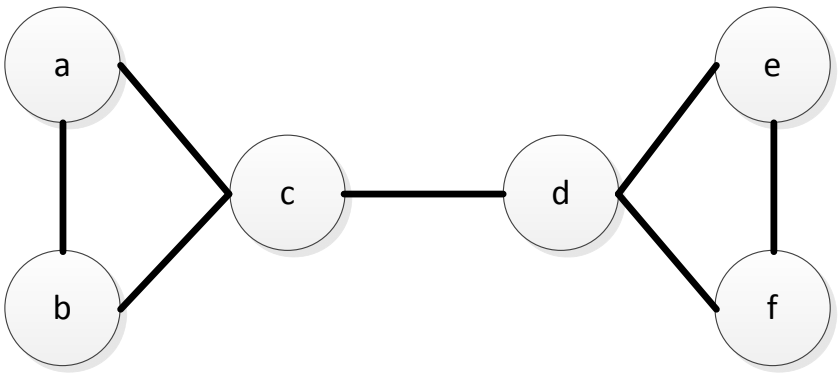

图 2.2 一个具有六个节点的网络结构图G^*。

Fig 2.2 A network G^* with 6 nodes.

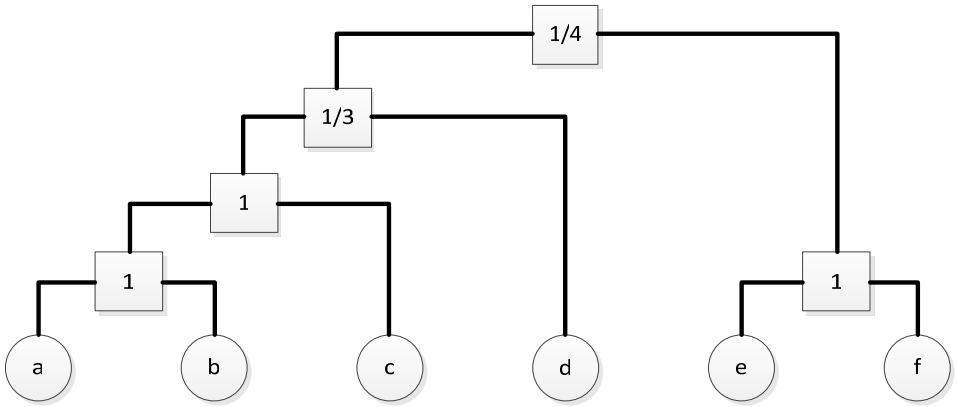

图 2.3 网络结构图G°的树状图D_1。

Fig 2.3 The dendrogram D_1 of the network G°.

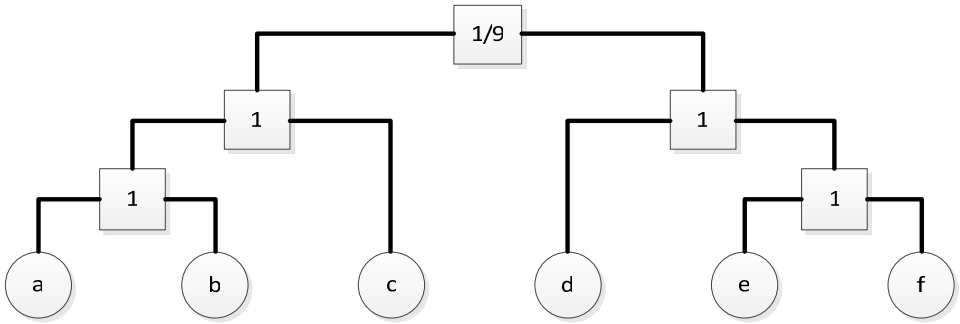

图 2.4 网络结构图G°的树状图D_2。

Fig 2.4 The dendrogram D_2 of the network G°.

对于一个固定树状图 D，可以容易求得

$$p_r^* = \frac{E_r}{L_r R_r} \tag{2.15}$$

因此，根据极大似然法，得到一个非常适合实际网络 G 的固定树状图 D，我们根据公式(2.15)可以很容易地确定参数集$\{p_r\}$。在图 2.2 中，给出了一个实际网络G°；在图 2.3 和 2.4 中，给出了两个树状图以及内置节点的概率。从形状上，我们就可以判断D_2比D_1更适合G°蕴涵的层次结构。在求树状图时，我们根据概率分布采用 MCMC 方法进行大量抽样来获得树状图。

在利用该模型进行链接预测时，要执行如下步骤。

① 抽样大量树状图并计算了所有内置节点的概率值。

② 对于任一对没有链接的节点x和y，我们根据所有的抽样样本的p_{xy}，计算平均概率值$\langle p_{xy}\rangle$。

③ 对$\langle p_{xy}\rangle$进行降序排列，概率较大的就是发生链接可能性大的节点对。

根据该模型在恐怖组织网络中实验的 AUC 值，我们可以得出该模型在链接预测领域要优于共同邻居方法。

第二类，基于概率关系模型的链接预测方法。

概率关系模型主要是将整个网络结构抽象为一个概率模型，然后，通过数据集学习各种参数，最后在学到的模型上进行链接预测。该模型主要分为两类，一是关系贝叶斯网(Relational Bayesian Networks: RBNs[34])、二是关系马尔科夫网(Relational Markov Networks: RMNs[35])。

概率关系模型 PRMs 是一个关系数据集上属性的联合概率分布。PRMs 允许对象的属性概率依赖于其他属性和其他对象的属性。该模型与基于单表的传统数据挖掘方法的最大不同是，可以在多关系数据集上进行挖掘。一个 PRMs 模型主要包括三部分：关系数据集G_D、关系依赖图G_M和推理图G_I。

关系数据集$G_D = (V_D, E_D)$，展示了 PRMs 模型的输入数据集组织结构，图中的节点是数据集中的对象，图中的边是对象与对象之间的关系。在G_D中，每个节点$v_i \in V_D$，每条边$e_j \in E_D$，并且$T(v_i) = t_{v_i}$和$T(e_j) = t_{e_j}$。每项都有$t \in T$，X^t表示类型t的所有属性。因此每个对象节点v_i的t_v类型和边e_j的t_e类型，都有$x_{v_i}^{t_{v_i}}$和$x_{e_j}^{t_{e_j}}$的属性集合。所以 PRMs 的联合概率分布是建立在

$$x = \left\{ x_{v_i}^{t_{v_i}} : v_i \in V_D, T(v_i) = t_{v_i} \right\} \cup \left\{ x_{e_j}^{t_{e_j}} : e_j \in E_D, T(e_j) = t_{e_j} \right\} \tag{2.16}$$

关系依赖图$G_M = (V_M, E_M)$，展示了G_D中的每项数据所属类型下的属性之间的依赖关系。每项数据的属性可以概率依赖该项数据的其他属性，以及其他相关对象或链接的属性。G_M中的每个节点V_M表示，$X_j^t \in X^t : t \in T$中的每一个属性。这些属性在G_D中有相同的类型，被联系在一起，因此每个类型在G_D中，根据类型属性之间的依赖关系可以划分成多个实例。在G_M中，主要包括两部分，一是所有类型属性之间的依赖关系，二是所有类型属性的条件概率分布(Conditional probability distribution: CPD)。

推理图$G_I = (V_I, E_I)$，展示了一个测试集中所有变量之间的概率依赖关系。G_D和G_M决定了G_I的结构。

基于关系贝叶斯网 RBNs 的链接预测方法。

在该模型中，G_M是一个有向无环图和条件概率分布表，P表示了所有类型属性的联合概率分布。其中pa_x表示节点x的双亲节点。该模型的联合概率表示如下式。

$$p(x) = \prod_{t \in T} \prod_{X_i^t \in X^t} \prod_{v:T(v)=t} p\left(x_{v_i}^t \mid pa_{x_{v_i}^t} \right) \prod_{e:T(e)=t} p\left(x_{e_j}^t \mid pa_{x_{e_j}^t} \right) \tag{2.17}$$

我们在构造该模型的依赖关系图时，特别注意不能在依赖关系图中存在环，因为存在环，这个模型就是无意义的。该模型需要对每个类型的每个属

性构造一个完整的条件分布表，这在大型复杂网络中是很难办到的。

基于关系马尔科夫网 RMNs 的链接预测方法。

一个 RMN 用一个无向图和势能函数ϕ来计算所有类型的所有属性的联合概率分布。C表示 RMN 中的 clique 集合，任一个 clique $c \in C$，c 包括了构成 clique 的$X_c (\in X^c)$变量集合。势能函数$\phi_c(x_c)$是一个非负的实数，其中$x_c \in X_c$。该模型的联合概率分布定义如下公式。

$$p(x)=\frac{1}{Z}\prod_{c\in C}\phi_c(x_c) \tag{2.18}$$

其中$Z = \sum_{x\in X}\prod_{c\in C}\phi_c(x_c)$是一个归一化常量。该类算法的学习过程有一个很高的计算复杂度。

2.2 马尔科夫逻辑网

2.2.1 马尔科夫逻辑网基础

如何将概率方法和一阶逻辑方法结合成为一种形式是人工智能研究的一个目标。P. Domingos 和 M. Richardson[36]提出了一种将概率方法和一阶逻辑方法结合起来的简单模型。概率模型能够有效地处理知识中的不确定性。一阶逻辑可以紧凑地表达广泛领域的知识。但是，在现实应用中，我们需要两个模型的功能。以前，提出的许多模型都过于复杂，并且适用条件苛刻。在本文中，我们将介绍 P. Domingos 和 M. Richardson[36]提出的概率结合逻辑的马尔科夫逻辑网模型(Markov Logic Network: MLN)，该模型仅要求数据的定义域是有限的。在 MLN 中，可以高效地学习和推理。

MLN 是一个每条逻辑子句都带有权重的逻辑知识库，是构造马尔科夫网的一个模板。在概率的观点上，MLN 定义了一个非常大的马尔科夫网，并且有能力表达广泛的领域知识。在一阶逻辑的观点上，MLN 可以处理不确定、不完整和有矛盾的知识。统计关系领域中的许多研究问题，可以直接地转化到 MLN 中研究。

在 MLN 模型中，主要用到了马尔科夫网和一阶逻辑的基础知识。

首先，给出马尔科夫网的简单介绍。

马尔科夫网也被称为马尔科夫场，是计算随机变量$X = (X_1, \cdots, X_n) \in \aleph$联合概率的模型。它是由一个无向图$G$和一系列势能函数$\phi_k$组成。在这个图$G$中，每个节点代表一个随机变量，每个团(clique)对应一个势能函数ϕ_k。一个势能函数在团(clique)的每一状态的取值为一个非负实数值。马尔科夫网定义的联合概率模型如下式。

$$P(X=x)=\frac{1}{Z}\prod_{k}\phi_k\left(x_{\{k\}}\right) \tag{2.19}$$

其中，$x_{\{k\}}$是第k个 clique 的状态，随机变量在 clique 中的取值。其中，$Z=\sum_{x\in X}\prod_k\phi_k\left(x_{\{k\}}\right)$是一个归一化常量。马尔科夫网，经常被转换为对数线性模型，其中每个 clique 的势能函数，被替换为对该 clique 状态的特征函数进行加权求和的指数形式。定义如下式。

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_{j}w_j f_j(x)\right) \tag{2.20}$$

一个特征函数可能是状态的任一实数值函数。在 MLN 中，主要采用了只有两个取值的特征函数，$f_j(x)\in\{0,1\}$。我们如果将势能函数ϕ_k直接转换到公式(2.20)中，每个 clique 的$x_{\{k\}}$一个状态都对应着一个特征函数，并且它的权重为$log\phi_k\left(x_{\{k\}}\right)$。根据 clique 的大小，这种转换方式呈指数级增长，所以尤其是当一个 clique 很大时，我们一般都是定义很小数量的，比势能函数表达更紧凑的特征函数。在 MLN 模型中，我们就利用了这点。

其次，对 MLN 中用到的基础知识一阶逻辑进行简单介绍。

一阶逻辑知识库(KB)，就是逻辑子句或公式的集合。公式主要是有常量、变量、函数和谓词构成。常量代表某一领域中的一个对象，如：在人的范围中，张三，李四等。变量可以取值某一领域中的任一对象。函数主要是将一组对象映射到另一对象上，如 MotherOf。谓词表示对象或对象属性之间的关系。

项(term)：表示任一领域中一个对象的表达式。该表达式可以是一个常量、变量或函数。例如：Anna，x，GreatestCommonDivisor(x, y)都是项。原子公式是一组项，应用于一谓词上构成。例如：Friends(x, Motherof(Anna))。公式是使用逻辑连接符和量词于原子公式递归构成的。如果F_1和F_2是公式，下面的也是公式。①$\neg F_1$为真，当且仅当F_1为假。②$F_1\wedge F_2$为真，当且仅当F_1和F_2都为真。③$F_1\vee F_2$为真，当且仅当F_1或F_2为真。④$F_1\Rightarrow F_2$(蕴涵)为真，当且仅当F_1假或F_2真。⑤$F_1\Leftrightarrow F_2$(等价)为真，当且仅当F_1和F_2有共同的取值。⑥$\forall xF_1$(全称量词)为真，当且仅当对于定义域中的每个x，F_1都为真。⑦$\exists xF_1$(存在量词)为真，当且仅当在定义域中至少有一个x，使得F_1为真。在公式的构造过程中，也可以添加括号改变逻辑连接符和量词的辖域范围。一个正实例是原子公式，一个负实例是一个否定原子公式。在 KB 中的公式隐含为是合取的，所以 KB 可以看作只含有一个大公式。一个实例化的项是，项中没有变量。一个实例化的原子公式或谓词是原子公式中的所有项是已经实例化的项。一个可能的

世界是对每个实例化的原子公式赋真假值。

一个公式是可满足的，当且仅当存在至少一个世界是它真。在一阶逻辑中，最基本的推理问题是，KB 是否蕴涵一个公式F($KB \models F$)。为了可以自动推理，我们经常把公式转化成一个合取范式。在 KB 中，公式是合取的，而公式内部的子句是析取的。在一阶逻辑中，所有公式都可以转化成合取范式。

在一阶逻辑中的推理是半自动的。由于这个限制，我们经常使用一阶逻辑的一个具有更多限制的子集，构造知识库。Horn 子句就是，最常用于构造知识库的子集，在一个 Horn 子句中，最多存在一个正实例。

在下面的表 2.3 中，展示了一个简单的知识库，并且和它的合取范式子句。我们可以看到这些公式在绝大多数情况下，总是成立的。但是，在现实应用中，我们很难找到一个总是成立的公式。因此，尽管一阶逻辑有强大的知识表示能力，但是纯粹的一阶逻辑很难应用于人工智能的实际问题。

在表 2.3 中，Fr()是Friends()的缩写，Sm()是Smokes()的缩写，Ca()是Cancer()的缩写。

表 2.3 一阶逻辑知识库和马尔科夫逻辑网的例子。

Table 2.3 Example of a first-order knowledge base and MLN.

中文	一阶逻辑	范式	权重
朋友的朋友是朋友	$\forall x\forall y\forall z Fr(x,y)\wedge Fr(y,z)\Rightarrow Fr(x,z)$	$\neg Fr(x,y)\vee\neg Fr(y,z)\vee Fr(x,z)$	0.7
没有朋友的人吸烟	$\forall x(\neg(\exists y Fr(x,y)))\Rightarrow Sm(x)$	$Fr(x,g(x))\vee Sm(x)$	2.3
吸烟致癌	$\forall x Sm(x)\Rightarrow Ca(x)$	$\neg Sm(x)\vee Ca(x)$	1.5
两人是朋友，都吸烟或否	$\forall x\forall y Fr(x,y)\Rightarrow(Sm(x)\Leftrightarrow Sm(y))$	$\neg Fr(x,y)\vee Sm(x)\vee\neg Sm(y)$ $\neg Fr(x,y)\vee\neg Sm(x)\vee Sm(y)$	1.1 1.1

最后，对马尔科夫逻辑网(MLN)进行介绍。

我们可以将一阶逻辑知识库，看作是可能世界集合上的一系列硬规则。如果一个世界违反了甚至一个规则，它的概率也是零。但 MLN 的基本思想是，软化这些规则，即使当一个世界违反了 KB 中的一个子句，它发生也是可能的，而不是不可能的。一个世界违反 KB 中的规则越少，发生的可能性就越大。每个公式或子句都有一个权重，体现了它的影响有多强。

定义 2.3 MLN：一个马尔科夫逻辑网L是(F_i, w_i)的集合，其中F_i是一阶逻辑子句，w_i是一个实数。并且结合一个有限的$C=\{c_1,\cdots,c_{|C|}\}$常量集合，这就

定义了一个马尔科夫网$M_{L,C}$如下。

① $M_{L,C}$对于每一个在L中出现的实例化谓词，都包含一个对应的二值节点。当实例化的原子公式为真时，这个二值节点取值 1，否则取值为 0。

② $M_{L,C}$对于每一个在L中实例化的逻辑子句，都包含一个对应的特征函数。当这个实例化的逻辑子句为真时，这个特征函数取值为 1，否则为 0。这个特征函数的权重为w_i对应于L中的F_i的权重。

逻辑子句在 MLN 中是标准的一阶逻辑语法，自由变量被看作逻辑子句最外层的全称量词。一个 MLN 可以看作一个构造马尔科夫网的模板，对于不同的常量集合，它将产生不同的马尔科夫网，并且马尔科夫网的规模可能有很大变化，但是其中的 clique、参数和特征函数等是不变的。根据公式(2.19)和公式(2.20)，$M_{L,C}$的联合概率分布定义如下公式(2.21)。

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right)=\frac{1}{Z}\prod_i \phi_i\left(x_{\{i\}}\right)^{n_i(x)} \tag{2.21}$$

其中，$n_i(x)$是实例化的F_i在x中成立的数目，$x_{\{i\}}$是F_i中的原子公式的值，并且$\phi_i(x_{\{i\}})=e^{w_i}$。该公式展示了将概率方法与逻辑知识融入一个模型的简便策略。

$M_{L,C}$的图形结构遵循定义 2.3，$M_{L,C}$的两个节点之间有边存在，当且仅当这两个实例化的原子公式一起出现在L的至少一个逻辑子句中。因此，在$M_{L,C}$中，L中的每个实例化逻辑子句中的原子构成一个 clique。

假设，存在常量 Anna 和 Bob，我们利用表 2.3 中的最后两个逻辑子句构造一个马尔科夫网$M_{L,C}$，如下图 2.5 所示。

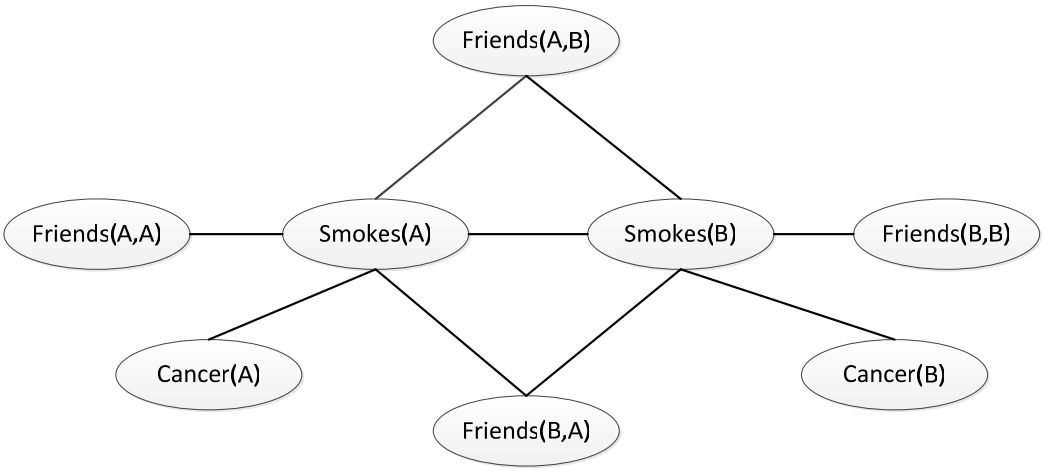

图 2.5 将常量 Anna(A)和 Bob(B)应用于表 2.3 中最后两个公式获得实例化的马尔科夫网。

Fig 2.5 Ground Markov network obtained by applying the last two formulas in Table 2.3 to the constants Anna(A) and Bob(B).

在图 2.5 中，如果每对节点一起出现在表 2.3 中最后两个实例化公式中，则它们之间有边相连。

图 2.5 的$M_{L,C}$中的每个状态代表一个可能世界，一个可能世界是对象、函数和对象之间关系的集合，并且有一个解释，决定每个实例化原子公式的真假值。

如果 Anna 和 Bob 有着吸烟的共同爱好。根据图 2.5 和公式(2.21)，计算他们是朋友的概率计算式如下：

$$
\begin{aligned}
&P\left(Fr\left(A,B\right)\mid Sm\left(A\right),Sm\left(B\right)\right)\\
&=\frac{P\left(Fr\left(A,B\right),Sm\left(A\right),Sm\left(B\right)\right)}{P\left(Sm\left(A\right),Sm\left(B\right)\right)}\\
&=\frac{P\left(Fr\left(A,B\right),Sm\left(A\right),Sm\left(B\right)\right)}{P\left(Fr\left(A,B\right),Sm\left(A\right),Sm\left(B\right)\right)+P\left(\neg Fr\left(A,B\right),Sm\left(A\right),Sm\left(B\right)\right)}\\
&=\frac{\frac{1}{Z}\exp\left(1.1\times 2\right)}{\frac{1}{Z}\exp\left(1.1\times 2\right)+\frac{1}{Z}\exp\left(1.1\times 2\right)}\\
&=0.5
\end{aligned}
$$

可知 Anna 和 Bob 在有吸烟的共同爱好下，是朋友的概率为 0.5。

如果 Anna 和 Bob 是朋友，并且 Bob 患有癌症和 Anna 吸烟。根据图 2.5 和公式(2.21)，Anna 患有癌症的概率计算式如下所示。

$$
\begin{aligned}
&P\left(Ca\left(A\right)\mid Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)\\
&=\frac{P\left(Ca\left(A\right),Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)}{P\left(Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)}\\
&=\frac{P\left(Ca\left(A\right),Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)}{P\left(Ca\left(A\right),Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)+P\left(\neg Ca\left(A\right),Fr\left(A,B\right),Sm\left(A\right),Ca\left(B\right)\right)}\\
&=\frac{\frac{1}{Z}\exp\left(1.5\times 2+1.1\times 1\right)}{\frac{1}{Z}\exp\left(1.5\times 2+1.1\times 1\right)+\frac{1}{Z}\exp\left(1.5\times 1+1.1\times 1\right)}\\
&=0.81
\end{aligned}
$$

我们可以计算 Anna 患有癌症的概率为 0.81。

2.2.2 马尔科夫逻辑网推理

MLNs 可以回答任意形式的条件概率，如：“在F_2的条件下，F_1的概率是多少？”如果F_1和F_2是一阶逻辑子句，C是一个出现在F_1或F_2中的有限常量集合，并且L是一个 MLN。

$$
\begin{aligned}
P(F_1 \mid F_2, L, C) &= P(F_1 \mid F_2, M_{L,C}) \\
&= \frac{P(F_1 \wedge F_2 \mid M_{L,C})}{P(F_2 \mid M_{L,C})} \\
&= \frac{\sum_{x \in X_{F_1 \cap F_2}} P(X = x \mid M_{L,C})}{\sum_{x \in X_{F_2}} P(X = x \mid M_{L,C})}
\end{aligned} \tag{2.22}
$$

其中，X_{F_i}是满足一阶逻辑子句F_i世界的集合。$P(X = x|M_{L,C})$在公式(2.21)中给出了定义。在马尔科夫网中，一般的条件概率查询是公式(2.22)的特例，这里所有在F_1、F_2和L中的谓词都是零维的，并且所有的公式都是合取的。

一个知识库 KB 在一阶逻辑中是否蕴涵一个公式F的问题，等价于是否$P(F|L_{KB}, C_{KB,F}) = 1$，其中，$L_{KB}$是将 KB 的所有逻辑子句的权重赋值为无穷大的 MLN，$C_{KB,F}$是出现在 KB 或F中所有的常量。当F为真，KB 是否蕴涵F，根据公式(2.22)计算$P(F|L_{KB}, C_{KB,F})$的概率可得。

直接计算公式(2.22)，即使在很小的定义域中，也是不可取的。由于 MLN 推理包含有#P-完全问题的概率推理和具有 NP-完全问题的逻辑推理，甚至是在一个很小的定义域中，也很难得到好的结果。

然而，由于大量高效的推理算法应用于了 MLNs，MLNs 允许精炼的知识编码，并且具有在特定上下文独立的特性，所以在一些例子中，MLNs 的推理效率要高于在同样定义域中，一般的图模型。从逻辑的观点上看，伴随高效的特征函数计算方式，MLNs 的概率意义使得近似推理变得更加容易。

在计算$P(F_1|F_2, L, C)$的概率时，一般的做法是在 MCMC 算法中，阻止进入F_2不成立的任何状态，然后计算抽得样本中F_1成立的数目，近似求得它的概率。然而，对于计算任何公式的概率，这种做法是非常耗时的。

所以在 MLNs 中对于F_1和F_2都是合取公式，提出了一种新的推理算法。尽管，该算法没有公式(2.22)通用，但是，该算法的语法是实际中经常用到的，并且该算法的执行效率远高于直接使用公式(2.22)。该算法主要分两个步骤，类似于构建知识库模型。

第一步，构建计算$P(F_1|F_2, L, C)$的极小实例化的马尔科夫网，构造过程如表 2.4 算法所示。

表 2.4 在 MLNs 中的推理网络构造算法

Table 2.4 Network construction for inference in MLNs

function ConstructNetwork(F_1, F_2, L, C)
input: F_1，不知真假值的实例化原子公式集合(“查询”)

F_2，知道真假值的实例化原子公式集合(“证据”)
L，一个马尔科夫逻辑网
C，常量集合
Output: M，一个实例化马尔科夫网
Calls: $MB(q)$，q在$M_{L,C}$的马尔科夫毯
$G \leftarrow F_1$
while $F_1 \neq \phi$
for all $q \in F_1$
if $q \notin F_2$
$F_1 \leftarrow F_1 \cup (MB(q) \backslash G)$
$G \leftarrow G \cup MB(q)$
$F_1 \leftarrow F_1 \backslash \{q\}$
return M，由G中的节点组成的实例化马尔科夫网，其中所有的边以及clique 和权值都对应于$M_{L,C}$。

ConstructNetwork 算法返回比原马尔科夫网$M_{L,C}$规模小的极小实例化马尔科夫网，由于凡是逻辑子句被F_2蕴涵，则可以忽略该子句并且相应的边也从M中去掉，使得推理算法执行效率提高。在最坏的情况下，该网络构造算法的空间复杂度为$O(|C|^a)$，其中a是谓词的最大维数，但是在实际中，谓词的维数通常很小。

第二步，将F_2赋值于M，并且在M上执行推理。在 MLN 中，通常采用 Gibbs 抽样方法，但是任何推理算法都可以使用。Gibbs 抽样方法的基本步骤是考虑到马尔科夫毯，抽样实例化原子公式。

定义 2.4 马尔科夫毯：是实例化的原子公式集合出现在，一个原子公式所属的公式集合。

当马尔科夫毯B_l在b_l状态时，实例化原子公式X_l的概率计算公式如下式。

$$P(X_l = x \mid B_l = b_l) = \frac{\exp\left(\sum_{f_i \in F_l} w_i f_i(X_l = x, B_l = b_l)\right)}{\exp\left(\sum_{f_i \in F_l} w_i f_i(X_l = 0, B_l = b_l)\right) + \exp\left(\sum_{f_i \in F_l} w_i f_i(X_l = 1, B_l = b_l)\right)} \quad (2.23)$$

其中，F_l是实例化公式中出现原子公式X_l的集合，当$X_l = x$和$B_l = b_l$时，$f_i(X_l = x, B_l = b_l)$是对应于第i个实例化公式，取值为 0 或 1 的特征函数。$P(F_1 \mid F_2, L, C)$的概率估计就是F_1中为真的合取公式，在整个抽样样本中所占的比例。

2.2.3 马尔科夫逻辑网学习

MLN 权重是通过一个或多个关系数据库学习获得。在学习的过程中，我们有一个“封闭世界假设”，如果一个原子公式不存在数据库中，则认为该原

子公式的取值为假。如果有n个原子公式，一个数据库可以表示成向量$x=(x_1,\cdots,x_i,\cdots x_n)$，其中$x_i$是第$i$个原子公式，如果该原子公式出现在数据库中，则$x_i=1$，否则$x_i=0$。原则上，MLN 权重可以通过标准方法在一个数据库中学得。如果，在x中，有$n_i(x)$个第i公式为真，对公式(2.21)求导得对应公式权重的导数如下。

$$\frac{\partial}{\partial w_i}\log P_w(X=x)=n_i(x)-\sum_{x'}P_w(X=x')n_i(x') \tag{2.24}$$

其中的求和公式是对所有世界进行求和，并且$P_w(X=x')$就是$P(X=x')$用当前权重向量$w=(w_1,\cdots,w_i,\cdots)$求得。公式(2.24)给出了第$i$公式在数据库中成立的数目和根据当前模型求得期望之间的差别。

计算一个公式在数据库中为真的次数，通常是不可取的。所以，在大定义域中，通过对一个公式进行抽样并且检查是否在抽样样本中是否满足，近似求得公式在数据库中为真的数目。然而，在小定义域中和本文后边的实验中，我们采用了一种高效的算法计算公式满足的数目。

这种算法被广泛地用于空间统计、社交网络模型和语言处理等领域。

$$P_w^*(X=x)=\prod_{l=1}^{n}P_w(X_l=x_l\mid MB_x(X_l)) \tag{2.25}$$

其中，$MB_x(X_l)$是X_l在数据库中的马尔科夫毯。对公式(2.25)进行求导，如下式。

$$\frac{\partial}{\partial w_i}\log P_w^*(X=x)$$

$$=\sum_{l=1}^{n}[n_i(x)-P_w(X_l=0\mid MB_x(X_l))n_i(x[X_l=0])-P_w(X_l=1\mid MB_x(X_l))n_i(x[X_l=1])]$$

其中，$n_i(x_{[X_i=0]})$是第i公式在$X_i=0$，其他数据不变情况下，为真的数目，同样对于$n_i(x_{[X_i=1]})$。当计算公式(2.25)的求导式时，不需要在整个模型上进行推理。

2.3 本章小结

本章主要详细介绍了有关链接预测研究的理论技术。首先，我们对目前链接预测研究领域中的链接预测方法进行了全面的介绍、分析和对比，并且详细描述了各个链接预测方法的特征。最后，我们对马尔科夫逻辑网的研究领域给出了详细的介绍，定义了马尔科夫逻辑网，介绍了马尔科夫逻辑网的推理过程和学习过程。

3 时变网络及传统链接预测方法的研究分析

目前，在链接预测的研究领域中，不同的链接预测方法有着不同的应用背景，并且不同的链接预测方法在同一应用背景下的链接预测准确率也相差很大。

在本文的研究中，针对链接预测方法的不同应用背景， 我们选择了时变网络作为一个统一的应用背景，并且给出了时变网络的详细定义。在时变网络的统一背景下，我们对不同链接预测方法的特点进行了详细的分析研究。

3.1 时变网络

在时变网络中，对象与对象之间在不同的时刻都会有事件发生。在实际生活中，交友网络和校友网络等社交网络，作者之间合著关系组成的网络和购物网络等网络，都是时变网络的一个实例。社交网络分析、购物篮数据挖掘都是传统的研究领域，本文主要在时变网络上进行链接预测的研究。在定义 3.1 中，给出了时变网络的准确定义。

定义 3.1 时变网络 ζ：是 (V,E_i,τ_i) 组成的集合。其中，V 是对象的集合，$E_i=\{(v_i,v_j)\,|\,v_i,v_j\in V\}$，$E_i$ 是 τ_i 时发生的事件集合，$\tau_i\in(t_0,...,t_n)$。

在图 3.1 中，给出了时间序列 $(t_0,...,t_9)$ 上，发生的所有事件构成的一个具体时变网络 ζ^o。

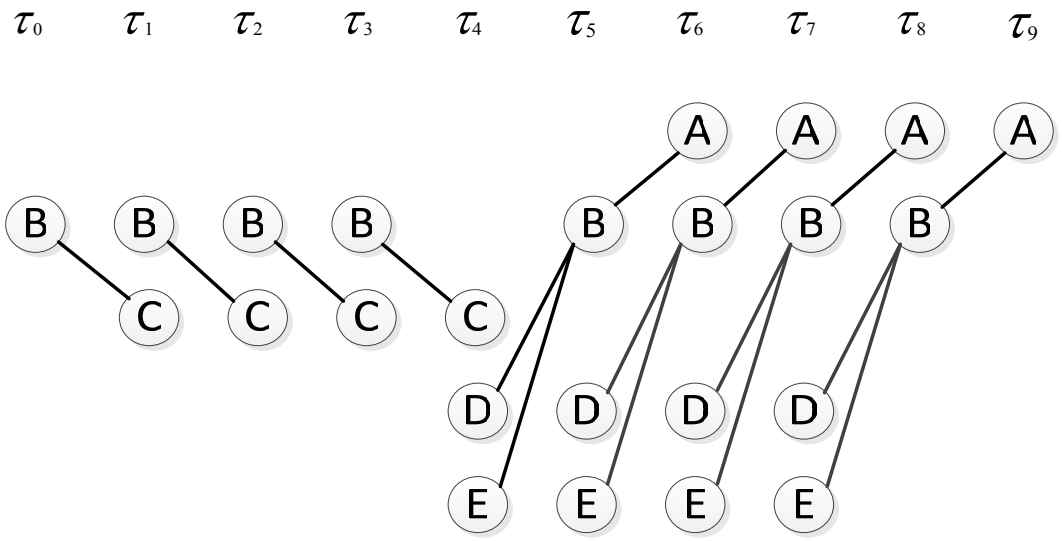

图 3.1 时变网络

Fig 3.1 Time-varying network

在时变网络上的链接预测的研究，一般都是基于时变网络的静态模型，如图 3.2 所示，在本文中，该模型称为时变网络静态模型。由 2.1.2 节可知，链接预测算法基本上都是基于时变网络的静态图模型[37]。

然而，在实际生活中，两个对象之间的关系是动态变化的，为了准确刻画两个对象之间关系的变化程度，本文提出了针对对象之间关系的“势”的概念来准确量化对象之间关系的变化。

图 3.3 是时变网络ζ^{o}的动态模型。根据图 3.2 和 3.3，我们可以看出时变网络静态模型，有如下缺点。

① 时变网络静态模型不能准确地反应对象之间的关系变化程度。

② 对于有时间序列要求的应用，时变网络静态模型不适用。

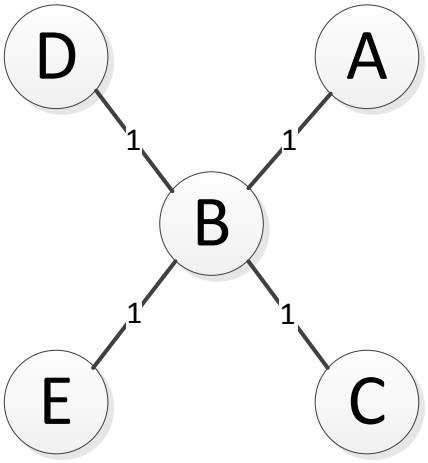

图 3.2 时变网络静态图

Fig 3.2 Static graph of the time-varying network.

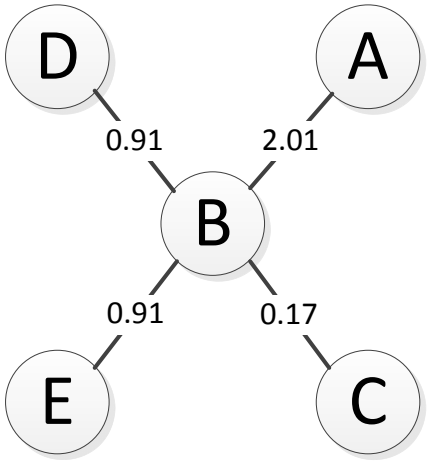

图 3.3 时变网络动态图

Fig 3.3 Dynamic graph of the time-varying network.

图 3.3 明显比图 3.2，描述两个对象之间关系的变化程度更加准确。因为在图 3.2 中，只是简单地统计两个对象之间是否有事件发生，而没有考虑事件发生的时间序列。但是，在链接预测方法研究中，事件发生的次数和时间序列都会对链接预测的准确率有影响。

3.2 链接预测方法研究分析

在社交网络的链接预测问题研究中，D. Liben-Nowell 和 J. Kleinberg 系统地比较了各种链接预测方法。

基于亲近度的链接预测算法，它们基本的流程如图 3.4 所示。

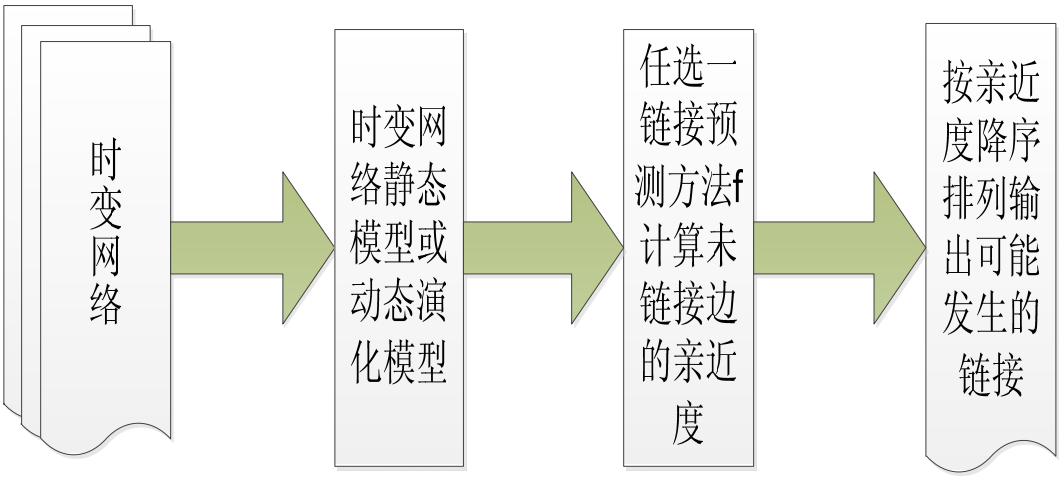

图 3.4 链接预测算法的基本流程。

Fig 3.4 The basic process of algorithms of link prediction.

最短路径链接预测方法遵循所有的交互网络都是“小世界”，在其中，对象之间是通过最短路径相互关联起来的[38]。该方法的亲近度是由对象之间的最短路径长度的负值决定的，我们将未发生链接的关系的亲近度，按照降序排列来进行链接预测。该类方法没有考虑对象的共同邻居数也没有考虑对象之间关系的变化程度，但是该方法在一些有局部关系网络构成的时变网络中是非常有效的。

基于两个对象之间共同邻居数[5]的链接预测算法的亲近度计算公式主要有 Common neighbors、Jaccard's coefficient、Adamic/Adar 等[11,12]。这些方法在存在大量链接信息的时变网络中链接预测效果明显，然而在非常稀疏的网络中，该类算法的链接预测结果非常差。

Katz 和 RWR 等基于对象之间路径的链接预测方法，需要整个时变网络的拓扑结构进行链接预测，并且该类算法的链接预测准确率明显高于基于邻近对象类的算法。但是，基于路径集合的链接预测算法，却存在着明显的缺点。一，计算一个网络的所有节点的亲近度是非常耗时的，甚至对于大型复杂网络是不可行的，二，整个网络的拓扑结构难以获得。

基于概率关系模型的链接预测方法一般是，建立整个链接网络的统一概率模型。该模型属于统计关系模型的一种，在该模型中可以采用多种方法进行链接预测。链接预测的概率模型通常都是基于马尔科夫随机场(Markov random fields)[18]的。Lise Getoor[19]等使用概率来量化属性和链接之间相互作用的关系，并且提出引用的不确定性和存在的不确定性两种机制来表示在链接图上的概率分布，建立一个概率关系模型。这类方法目前，存在的主要问

题是，在模型的建立中许多链接的先验概率难以获得，进行预测时计算量大。

3.3 本章小结

在本章中，我们给出了时变网络的定义，并且介绍了时变网络的静态模型和动态模型。通过对比时变网络静态模型和动态模型，我们给出了描述对象之间关系变化程度的势的概念。最后，全面地分析了基于时变网络的链接预测方法的特点。

4 基于时变网络的链接预测方法研究

链接预测的主要问题就是基于对象的属性和其他观测到的链接，预测两个对象之间是否存在链接。目前的链接预测研究，主要可以分为两类，一是基于网络结构的链接预测方法，二是基于概率关系模型的链接预测方法，这种方法主要是抽象整个网络结构为一个概率模型，学习各种参数，然后进行链接预测。在目前链接预测的研究领域中，基本上所有的链接预测方法都是基于时变网络静态模型的。

由于，时变网络静态模型不能准确地反映对象之间的关系变化程度，并且不适用于有时间序列要求的链接预测研究，同时，各种不同的链接预测方法在不同的数据集中的预测准确率都各不相同，在不同的环境中也表现迥异。通过对时变网络和马尔科夫逻辑网的研究，在本节中，我们提出了能够准确量化对象之间关系重要性的时变网络动态演化模型和改进的链接预测算法，并且还提出了一个基于马尔科夫逻辑网的链接预测方法多融合的新链接预测方法。

4.1 时变网络的动态演化模型及改进的链接预测算法

4.1.1 时变网络的动态演化模型

在传统的信息处理研究领域中，各种信息处理技术都是应用于时间 t_0 到时间 t_n 时变网络的一个整体静态的数据集[39]，而没有把时变网络抽象成为一个随时间不断变化的动态网络来处理，并且各种研究都是基于“对象-中心”的研究。

在本文中，我们的研究是基于“关系-中心”的研究，并且使用“势”的概念来精确量化对象之间关系的重要性，紧密性等性质。

由于时变网络静态模型，存在明显的缺点，并且不能有效地提高链接预测的准确率，所以在本文中，提出了一种准确描述时变网络动态模型的方法。在定义 4.1 中，给出了时变网络的动态演化模型的定义。

定义 4.1 时变网络动态演化模型 χ：是 (E_i, f_i) 组成的集合。其中，E_i 是到 τ_i 时刻所有关系的集合，$\mathrm{E}_i = \bigcup_{j=0}^{i} E_j$，$f_i$ 是关系 $e_i = (v_x, v_y)$ 在 τ_i 时的“势”，f_i 是一个实数，任一 $e_i = (v_x, v_y)$ 第一次发生事件时，关系的初始势 f_i 都为一，并且 f_i 满足 $0 \le f_i \le |\mathrm{E}_i|$ 限制。

由于在社交网络分析、网页排序、链接挖掘等领域中存在着各种不同的时变网络动态模型，所以给出一个描述时变网络动态模型的一致性准则。根

据时变网络动态模型的特性，我们提出了如下几个准则，并且所有基于时变网络的模型都应该满足如下几个准则：

① 在时变网络的任一时刻中，对象与对象之间的关系都是可比的。

② 对象之间发生了事件，保证该对象之间关系的权值有所增加或至少不变。

③ 对象之间未发生事件，保证该对象之间关系的权值有所减少或至多不变。

④ 所有对象之间关系的权值和保持不变，即事件的重复发生不会引起时变网络的权值和增加或减少，关系的权值也不可以随意变化。

⑤ 事件序列和事件发生的次数对关系的权值有影响。

本文提出的时变网络动态演化模型(Time-Varying Network Model：TVNM)的主要思想：在 χ 中，两个对象之间的势越大，表示两个对象之间的关系越紧密、亲密和牢固等；反之，两个对象之间的关系越疏远。由于时变网络动态模型能够反映时间序列的变化，所以，对象之间的势是随着时间的变化而变化。

本文借助对象之间关系的势的概念描述 TVNM 的方法，在 τ_i 时 TVNM 势的变化是基于时间 τ_{i-1} 时 TVNM 的势，势的流动仅受制于对象之间是否发生事件。

势的一般流动形式：在 τ_i 时未发生事件的关系中，这些关系的部分势，流向 τ_i 时发生了事件的关系。在 τ_i 时发生了事件的关系，总能获得一些势，并且至少不会降低。在 τ_i 时没有发生事件的关系，则损失部分势。因此，TVNM 满足了准则②和③。

在 χ 中，如果是对象之间第一次发生事件，则该对象之间关系的势 f 初始值为一。这样保证了所有关系的势 f 的差别，不是因为赋初值的缘故，而是，由于对象之间发生的事件序列，或事件重复发生的次数的原因。这也确保了各个对象之间的关系是可比的，即该模型满足准则①。

在 τ_i 时的 χ_i 中，每个关系的势 f_i 是通过前一时刻 χ_{i-1} 定义，并且给出每个关系的势 f 的递归定义，如下公式(4.1)。

$$f_i(e)=\begin{cases} f_{i-1}(e)+\alpha_i\dfrac{\overline{f_{i-1}(e)}}{\sum_{d\in P_i}\overline{f_{i-1}(d)}} & e\in P_i \\ f_{i-1}(e)\left(1-\dfrac{\alpha_i}{T_{i-1}}\right) & e\notin P_i \end{cases} \tag{4.1}$$

其中，α_i 是，所有 $e\notin P_i$ 的关系集合贡献给 $e\in P_i$ 的关系集合的总势，描

述了关系新发生事件的支持度。

如果α_i的值越大，关系越晚发生事件，则它的势就越大，反之，关系的势减小，P_i表示ζ_i中E_i的集合，$\overline{f_{i-1}(e)}=|\mathrm{E}_i|-f_{i-1}(e)$，$T_{i-1}$表示$e \notin P_i$的关系集合的总势。

在χ_0中，任一关系的势f都为 0。在公式(4.1)中，由于对象之间新发生的事件的分配比例较大，所以可以分配较多的势。

因为，实践证明新出现的事件相对会重要，如果新出现的事件不重要，则该对象的势会在后面的时间序列中慢慢变小，这也符合了社会中实际存在的关系或事件的重要性变化过程。

$$T_{i-1}=\sum_{d\notin P_i} f_i(d) \tag{4.2}$$

$$0\le\alpha_i\le T_{i-1} \tag{4.3}$$

在 TVNM 中，α_i的取值是一个关键因素影响到了模型的表征能力，当α_i取值为 0 时，该模型退化成时变网络静态模型，反之，则该模型不考虑历史信息，可以影响到α_i因素主要有如下几个，P_i集合的大小、事件上次发生到现在消逝的时间等。

在本文中，提出了一种方法量化α_i与T_{i-1}之间的关系，α_i等价于在T_{i-1}中所占的比例，为$\eta(\eta\in(0,1))$，$\alpha_i=\eta\bullet T_{i-1}$。

定理 4.1：在时变网络ζ中，重复出现的事件不会引起时变网络动态模型χ总势的改变。该定理简称：势守恒定理。

证明：要证，重复出现的事件不会引起总势的改变，假设时间τ_i到τ_{i+1}的时变网络中没有第一次发生的事件，只需证明时变网络动态图的总势没有改变。时间τ_i到τ_{i+1}的时变网络中只存在重复发生的事件，根据公式(4.1)在时间τ_{i+1}的时变网络动态图的总势是：

$$\begin{aligned}\sum_{e\in \mathrm{E}_{i+1}} f_{i+1}(e)&=\sum_{e\in P_{i+1}}\left(f_i(e)+\alpha_{i+1}\cdot\frac{\overline{f_i(e)}}{\sum_{d\in P_{i+1}}\overline{f_i(d)}}\right)+\sum_{e\notin P_{i+1}} f_i(e)\cdot\left(1-\frac{\alpha_{i+1}}{T_i}\right)\\&=\sum_{e\in P_{i+1}} f_i(e)+\alpha_{i+1}+\sum_{e\notin P_{i+1}} f_i(e)-\alpha_{i+1}\\&=\sum_{e\in \mathrm{E}_{i+1}} f_i(e)\end{aligned}$$

□

该定理证明了时变网络动态演化模型不会因为重复发生的事件，而改变该模型总势的大小，即满足了准则④。该定理也保证了对象之间的关系是可比的，即保证时变网络动态演化模型满足准则①。由于定义时变网络动态演

化模型考虑了事件发生的序列，即该模型也满足准则⑤，同时，也满足了时变网络动态模型应该具有的所有性质。

4.1.2 动态演化模型算法及演化过程

由于，时变网络静态模型突出的缺点，不适合基于时变网络的链接预测领域的进一步研究，所以，在本文中提出了时变网络的动态演化模型。在表 4.1 中，我们给出时变网络的动态演化模型的算法描述。

表 4.1 时变网络动态演化模型的算法描述。

Table 4.1 Algorithm of the time-varying network model.

function ConstructTVNM(ζ，τ_0，τ_n)

input: ζ，时变网络

τ_0，开始时间

τ_n，结束时间

Output: χ，一个已经赋值的时变网络动态演化模型

Calls: $Potential(e,\tau_i)$，根据公式(4.1)计算，e在ζ中在时刻τ_i的势。

for *i* = 1 to n do

P = $\zeta_t.E_t$ // P 表示在时刻τ_i发生的事件。

T = $\chi_t.E_t - P$ // T 表示在时刻τ_i没有发生，但以前发生过的事件。

for each $p \in P$

if $p \notin \chi_{t-1}.E_{t-1}$

$\chi_t(p,1)$

$\chi_t(p,Potential(p,\tau_i))$ //计算p在τ_i时刻的势

end for each

for each $q \in T$

$\chi_t(q,Potential(q,\tau_i))$ //计算q在τ_i时刻的势

end for each

end for

return χ，由ζ中发生事件的关系组成的时变网络动态图，其中所有关系都对应着势的大小。

对于表 4.1 中的算法，进行算法分析以后，我们可得该算法的时间复杂度为$O(n^2)$，空间复杂度为$O(n)$。

根据表 4.1 中的时变网络动态演化模型的算法，以时变网络ζ^0为例，设定参数$\eta = 0.3$，得到图 3.3 的时变网络的动态演化模型。在表 4.2 中显示了动态模型的演化过程。

此处设置时变网络动态演化模型的参数$\eta = 0.3$，主要是为了说明时变网络中关系的一个演化过程。参数 η 的具体设置，需要考虑特定的应用背景。

表 4.2 关系的势随时间变化。

Table 4.2 The potential of a relation changing with time.

关系	τ_5	τ_6	τ_7	τ_8	τ_9
(A, B)	0	1.07	1.12	1.16	2.01
(B, C)	0.7	0.49	0.34	0.24	0.17
(B, D)	1.15	1.22	1.27	1.30	0.91
(B, E)	1.15	1.22	1.27	1.30	0.91
总势	3	4	4	4	4

我们可知时变网络的静态模型，只是简单统计了两个对象之间是否发生事件，而没有准确地描述两个对象之间关系的紧密程度。从表 4.2，可以看出对象之间关系的一个动态演化过程，特别是在时刻τ_9以后，可以明显地看出关系(A, B)要远比其它关系紧密。虽然，当η的取值不同时，关系(A, B)、 (B, C)、(B, D)和(B, E)的势会不一样，但是这四个关系的紧密程度是不会改变的。

我们可以得出这样的结论，在时变网络的动态演化模型上，做链接预测研究，将会有明显地提高。若在图 3.3 上做链接预测研究，我们可以明显地知道关系(A, E)和(A, D)很有可能发生事件。所以，时变网络的动态演化模型的表征能力是优于时变网络的静态模型，在第五章中，提供了更进一步的实验证明。

4.1.3 基于动态演化模型改进的链接预测算法

在第 2.1.2 节中，我们探讨了国内外链接预测研究领域中的主流方法。由于这些链接预测方法的提出背景都是基于时变网络静态模型的，在 3.1 节中我们对比了时变网络静态模型和时变网络的动态演化模型，得出时变网络的动态演化模型的表征能力明显优于时变网络静态模型，所以本文改进了其中的一些链接预测算法以便适应于时变网络的动态演化模型。

在时变网络静态模型中，共同邻居方法主要是计算两个对象之间的共同邻居数目，如果共同邻居数越多，则两个对象发生链接的可能性就越高。在时变网络的动态演化模型下，改进的链接预测方法不只是计算两个对象之间的共同邻居数，而是加入了对象之间关系变化的因素。

在本文中，通过对表 2.1 和表 2.2 的分析，我们主要对 Common Neighbors (CN)、Jaccard、Sorensen、Resource Allocation(RA)和 Katz 等预测效果好的链接预测方法进行改进。

① 第一种改进的共同邻居方法(I_CN1) 。

$$S_{xy}^{I\_CN1}=\prod_{z\in\Gamma(x)\cap\Gamma(y)}\left(f(x,z)+f(z,y)\right) \tag{4.4}$$

其中，z是对象x和y之间的共同邻居，$f(x,y)$表示对象x和y之间的势。

② 第二种改进的共同邻居方法(I_CN2)

$$S_{xy}^{I\_CN2}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\left(f(x,z)+f(z,y)\right) \tag{4.5}$$

该算法中，两个对象之间的亲近度是求相邻所有关系的势的和，如果两个对象的亲近度越大，则两个对象之间发生链接的可能性越大。

我们对 I_CN1 和 I_CN2 进行算法分析，可知需要计算大约n^2个关系的势，并且每个关系的势的计算需要大约n步完成，所以，这两个算法的时间复杂度为$O(n^3)$。

③ 改进的 Jaccard 方法(I_Jaccard)。

$$S_{xy}^{I\_Jaccard}=\frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)}\left(f(x,z)+f(z,y)\right)}{\left|\Gamma(x)\cup\Gamma(y)\right|} \tag{4.6}$$

Jaccard 系数通常用来表示x和y都有一个特征f和在x或y都有特征f中的概率，并且被广泛地应用于信息检索领域计算亲近度。在本文中，我们将该方法用于链接预测领域，计算对象之间的亲近度。

I_Jaccard 算法也需要计算大约n^2个关系的势，并且每个关系的势的计算需要大约n步完成，所以，这个算法的时间复杂度为$O(n^3)$。

④ 改进的 Sorensen 方法(I_Sorensen)。

$$S_{xy}^{I\_Sorensen}=\frac{2\times\sum_{z\in\Gamma(x)\cap\Gamma(y)}\left(f(x,z)+f(z,y)\right)}{\sum_{u\in\Gamma(x)}f(x,u)+\sum_{v\in\Gamma(y)}f(v,y)} \tag{4.7}$$

该方法类似于 Jaccard，只是与 Sorensen 方法提出的背景不同。Sorensen 方法主要被用于生物社区数据的分析研究中，但是在链接预测领域，该算法的链接预测效果也是非常好。

⑤ 改进的 Resource Allocation 方法(I_RA)。

$$S_{xy}^{I\_RA}=\frac{1}{\sum_{z\in\Gamma(x)\cap\Gamma(y)}\sum_{u\in\Gamma(z)}f(u,z)} \tag{4.8}$$

在复杂网络的资源动态分配中，该方法取得了非常好的效果。如果，节点x和y没有链接，节点x可以借助于节点y的共同邻居，将一些资源发送给节点y。在改进的 Resource Allocation 方法中，每个传递者拥有的资源是随时间在变化的，节点x会将自己的资源平均地分发给它的邻居节点。同样的理论，节点y可以从节点x和y之间获得资源。

I_RA 算法也需要计算大约n^2个关系的势，并且每个关系的势的时间复杂度为$O(n^2)$，所以该算法的时间复杂度为$O(n^4)$。

⑥ 改进的 Katz 链接预测方法(I_Katz)。

根据$\left|paths_{xy}^{<l>}\right|$的计算方式不同，Katz 方法主要分为不加劝的 Katz 链接预测方法和加权的 Katz 链接预测方法。

第一类是不加劝的 Katz 链接预测方法，任何节点x和y之间有长度为l的路径，则有$\left|paths_{xy}^{<l>}\right|=1$，节点$x$和$y$之间没有长度为$l$的路径，则$\left|paths_{xy}^{<l>}\right|=0$。

第二类加权的 Katz 链接预测方法，任何节点x和y之间有长度为l的路径，则，有$\left|paths_{xy}^{<l>}\right|$等于节点$x$和$y$之间有长度为$l$的路径的条数。

改进的链接预测算法，主要是针对加权重的 katz 方法。

$$\left|paths_{xy}^{<l>}\right|=\sum_{l^{o}\in\chi_{xy}^{l}}\sum_{(u,v)\in l^{o}}f(u,v) \tag{4.9}$$

$$S_{xy}^{I\_Katz}=\sum_{l=1}^{\infty}\beta^{l}\left|paths_{xy}^{<l>}\right| \tag{4.10}$$

其中，χ_{xy}^{l}表示时变网络的动态演化模型中的对象x到y路径长度为l关系的集合，如：$\chi_{xy}^{l}=\{(x,u),(u,v),\cdots,(w,y)\}$且$\left|\chi_{xy}^{l}\right|=l$，$\left|paths_{xy}^{<l>}\right|$为节点$x$到$y$的长度为$l$的路径的值，$\beta$是一个预测参数且$\beta>0$。如果$\beta$很小接近于零时，则该方法近似于共同邻居方法。

4.2 一种新的基于 Markov 逻辑网的链接预测方法

MLN 是一个每条逻辑子句都带有权重的逻辑知识库，是构造马尔科夫网的一个模板，并且成功地将概率方法和一阶逻辑方法结合成为一种新的简单模型。MLN 融合了概率模型能够有效地处理知识中的不确定性和一阶逻辑可以紧凑地表达广泛领域的知识的优点，摒弃了它们各自的缺点。

从概率的观点上，MLN 定义了一个非常大的马尔科夫网，并且有能力表达广泛领域的知识。在一阶逻辑的观点上，MLN 可以处理不确定、不完整和有矛盾的知识。

在链接预测领域中的各种链接预测方法，都具有不同的特点，在不同的环境中也表现迥异，现在随着 P. Domingos 和 M. Richardson[36]提出了甚至可以包含相互矛盾的知识的 MLN 模型，并且可以将统计关系领域中的许多研究问题，可以直接地转化到 MLN 中研究。在本章中，将各种不同的传统链接预测方法融入，我们新提出的基于马尔科夫逻辑网的链接预测方法中，并且将基于时变网络的动态演化模型改进的链接预测算法也应用于了马尔科夫逻辑网。

我们知道传统的链接预测方法各有不同的特点，在不同的环境中预测结果的好坏也有明显差异，并且都对应着不同的计算方法，难以融入一个模型

和综合利用各个链接预测算法的不同优点。

随着，马尔科夫逻辑网的提出，逻辑方法与概率方法的结合给我们提供了一个将不同的传统链接预测算法融入一个模型中的机会。所以在本文中基于马尔科夫逻辑网和亲近度的链接预测算法，我们提出了一个新的链接预测算法。在本文中，这个新的链接预测算法用符号 MLP(Method of Link Prediction based on MLNs: MLP)表示。

该算法主要分为如下三个步骤。

① 由于马尔科夫逻辑网是一个基于一阶逻辑的系统，我们将基于亲近度的链接预测算法的值域离散化。根据离散化的结果，构造一个一阶逻辑系统。将时变网络的数据预处理为适合该逻辑系统的谓词库。

我们用 L(x, y)表示对象 x 和 y 之间的链接关系，时变网络转化成为一个包含谓词 L 的谓词库。

基于亲近度的链接预测算法的离散化过程。

假设，存在逻辑知识库 KB 和任一基于亲近度的链接预测算法f，$f(x,y)$表示对象 x和y之间的亲近度，则链接预测算法f的离散化过程如下式。

$$KB = KB \bigcup \begin{cases} f_1(x,y) & 0 \le f(x,y) < n_1 \\ f_2(x,y) & n_1 \le f(x,y) < n_2 \\ \vdots & \vdots \\ f_n(x,y) & n_{n-1} \le f(x,y) < n_n \end{cases} \tag{4.11}$$

其中$N = max(f(x,y))$，并且$0 \le n_i \le N, 1 \le i \le n$。在链接预测算法的离散化过程中，需要确定$n_i$的具体值。

建立逻辑系统：

假设，存在一阶逻辑系统 FOKB，则 FOKB 的建立过程如下式。

$$FOKB = FOKB \bigcup \{ f_i(x,y) \Rightarrow L(x,y) \mid f_i(x,y) \in KB \} \tag{4.12}$$

对于逻辑知识库中的谓词，是构建马尔科夫逻辑网的基础模型逻辑系统。

② 建立马尔科夫逻辑网。根据建立的逻辑系统和得到的谓词库，学习每条逻辑子句的权重。

③ 使用学习到的马尔科夫逻辑网，进行计算所有未发生链接关系的概率。根据未发生链接关系的概率，对所有未发生链接的关系进行排序，进行链接预测。

该算法中，一阶逻辑系统 FOKB 的谓词都是二维的。在推理时，该算法的空间复杂度为$O(2^{|C|})$，可知随着常量的增加，空间会急剧增长，所以该算法采用了马尔科夫毯来降低空间复杂度。

该算法的时间复杂度也是很高的[40, 41]，在推理的过程中，采用了 MCMC

算法做近似概率计算。

4.3 本章小结

本章提出了时变网络的动态演化模型，并且给出了动态演化模型的算法描述和势守恒定理的证明，详细分析了动态演化模型的特性。在时变网络的动态演化模型的基础上，我们提出了改进的链接预测算法。

同时，本章也提出了，基于时变网络的马尔科夫逻辑网的新的链接预测方法，该方法是将多种传统的链接预测方法融于马尔科夫逻辑网成为一个新的链接预测方法。

5 实验及实验分析

在第 4 章中，我们给出了基于时变网络动态演化模型的改进链接预测算法和基于马尔科夫逻辑网的新的链接预测算法。在此基础上，我们将改进的链接预测算法和新的链接预测算法，通过在数据集 Enron 上进行 10 折交叉实验，并且用标准的度量方法准确率和 AUC 进行度量。最后对改进的链接预测算法和新的链接预测算法的实验结果进行了全面、客观的实验分析。

5.1 数据集与 MLN 实验环境介绍

我们采用了由 CALO(A Cognitive Assistant that Learns and Organizes)项目收集和整理的数据集，并且该数据集被广泛地用于分类、簇类和链接预测等领域。该数据集是从 1999 年 5 月 11 日到 2002 年 6 月 22 日的大约三年时间里，Enron 公司 151 位员工发送和接收的 250,000 封邮件组成的[42, 43]。

从 Enron 数据集邮件的组成结构上分析，存在一部分人拥有大量邮件，并且存在一些人给公司其他所有员工发送邮件从 1 封到 10000 封。

该数据集是公开的，我们可以从(http://www-2.cs.cmu.edu/~enron/)上自由下载用于科学研究。

在 Marc, Sumner 和 Pedro, Domingos 开源的 Alchemy 平台上，我们主要利用 learnwts 和 infer 两个工具做 MLN 的相关实验[44]。

5.2 实验设置和链接预测算法性能的度量标准

在该数据集中，一些员工发送或接收了公司之外客户的电子邮件，这些数据对于链接预测没有用处。在链接预测的研究中，我们通常只是考虑两个对象之间是否会发生链接或者是否丢失了链接，而不特定去考虑是那一个对象引发了事件。

但是在 Enron 数据集中，每个人发送或接收邮件都是有特定方向的，在本文中，我们泛化了两个员工之间的发送和接收邮件关系，而不考虑特定的方向。

为了满足实验要求，我们去掉邮件的接收者或发送者不是该公司员工的所有邮件。通过数据集的预处理，该数据集包含了公司员工之间发送或接收的大约 47，000 封邮件。

在实验中，员工自己之间的发送和接收邮件都不考虑，我们将公司员工之间的邮件关系抽象成为一个无向图$G(V,E)$,其中V是节点，E是节点之间的链

接。

符号U表示所有对象之间的链接，包含了$|V|(|V|-1)/2$个链接，其中$|V|$是集合V中包含的元素数目。

符号$U-E$表示，图G中不存在的链接。我们假设一些丢失的链接或将来会发生的链接都在集合$U-E$中，所以链接预测的任务就是找出这些链接。

通常情况下，我们不知道那些链接是丢失的链接，那些链接是将来发生的链接，否则就不需要做链接预测研究了[45]。因此，为了测试一个链接预测方法的准确率，我们通常将已经存在的链接E随机分成两部分。

E^T表示训练集，是已经知道的链接集合(T: Train)。

E^P表示测试集，是需要预测的链接集合(P: Probe)。

显然，有$E^P \cap E^T = \phi$和$E^P \cup E^T = E$成立。

随机抽样验证的优点是训练集和测试集的比例不依赖于迭代的次数，但是该验证方法的缺点是一些链接不会被抽为测试集中的一部分，而另一些链接可能多次被抽为测试集中的一部分，所以该方法会导致一些统计误差。

K 折交叉验证方法可以有效地克服随机抽样验证方法的缺点。该方法是将存在的链接集合E随机地分成 K 份儿，每一份当且仅当的只做一次测试集，其他 K-1 份作为训练集，验证过程迭代 K 次。很显然的，一个较大的 K 会导致较小的统计偏差，但是计算量会显著地增加。实验表明 10 折交叉验证是一个精度与计算量很好的平衡点[46]。

衡量链接预测算法性能的两个标准度量方法：

①AUC(area under the receiver operating characteristic curve)曲线。

②准确率度量方法。

一般情况下，对于所有未发生链接关系的集合($U-E^T$)，一个链接预测算法给出一个降序排列，或者对于其中的每一个未发生链接的关系给出亲近度值。AUC 是根据整个未发生链接关系的列表来度量链接预测算法，而准确率只是注重于前n个链接关系。

① AUC 度量标准

一个链接预测算法的 AUC 值可以被解释为概率，在测试集E^P中，随机挑选一个链接的概率要大于，在没有发生事件的链接$U-E$中，随机挑选一个链接的概率。

在对一个链接预测算法的实际度量过程中，我们通常只是计算每一个未发生链接关系的近似度值，而不是给出一个有序的列表，因为后者通常是很耗时的。

因此，每次我们随机挑选一个丢失的链接和未发生的链接，比较它们的

亲近度值，如果在n个独立的比较中，有n'次丢失链接的近似度大于未发生链接的近似度，有n''次丢失链接的近似度等于未发生链接的近似度，则这个算法的 AUC 值定义如下。

$$AUC=\frac{n'+0.5n''}{n} \tag{5.1}$$

如果所有未发生链接的近似度服从一个独立同分布的概率，则它的 AUC 值为 0.5。因此，一个链接预测算法的 AUC 值大于 0.5 的程度，表示了该算法比单纯猜测方法好的程度。

② 准确率

假设对所有未发生的链接进行排序后，在选定未发生链接的比例下，预测准确的链接所占的比例。

$$\text{准确率}=\frac{\text{检索出的相关信息量}}{\text{检索出的信息量}}\times 100\%$$

$$\text{检出率}=\frac{\text{检索出的信息量}}{\text{检索中的信息总量}}\times 100\%$$

例如，我们取前L个链接，在这个链接集合中，有L_r个链接预测准确，则准确率就是L_r/L，检出率就是$L/|U-E^T|$。很显然的，在检出率确定的情况下，准确率越高算法的就越好。

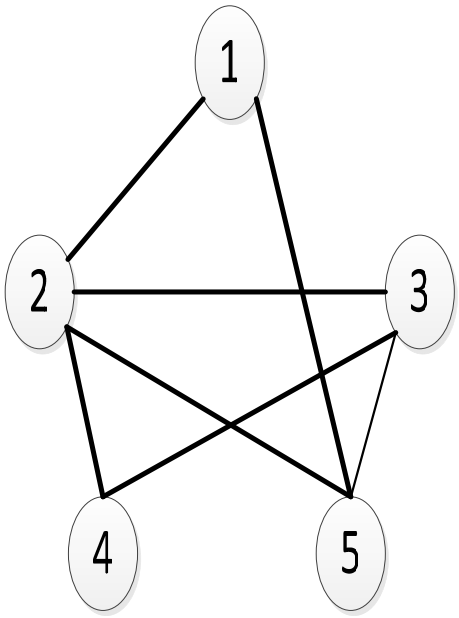

图 5.1 完整图

Fig 5.1 Whole graph

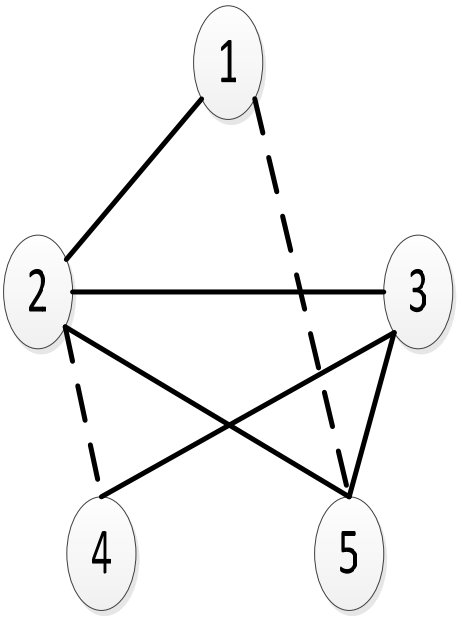

图 5.2 训练图

Fig 5.2 Training graph

图 5.1 和 5.2 作为例子，介绍了如何计算 AUC 和准确率。在图 5.1 的简

单图中，5 个节点，7 条存在的链接和 3 个不存在的链接((1, 3), (1, 4)和(4, 5))。为了测试一个链接预测算法的精度，我们需要将图 5.1 分成测试集和训练集，例如图 5.2 的实线部分组成一个训练集，虚线部分组成一个测试集。一个链接预测算法在进行链接预测时，只允许使用训练集的信息。

假设存在一个链接预测算法，将所有不存在链接的边都进行了亲近度计算，分别给予边$S_{13}=0.4$，$S_{15}=0.5$，$S_{14}=0.6$，$S_{45}=0.5$和$S_{24}=0.6$。

评估该算法的 AUC 值：从图 5.2 中，可知总共有 6 个未链接和测试集的比较对，分别是$S_{15}>S_{13}$，$S_{15}<S_{14}$，$S_{15}=S_{45}$，$S_{24}>S_{13}$，$S_{24}=S_{14}$和$S_{24}>S_{45}$。于是可得该算法的 AUC 值为(3×1＋2×0.5)/6≈0.67。

计算该算法的准确率：假设L等于 2，则亲近度最高的两个链接分别是(1, 4)和(2, 4)，其中(2, 4)属于测试集，因此该算法的准确率为 0.5。

5.3 基于动态演化模型的改进的链接预测算法的实验分析

将 Enron 实验数据集随机分成 10 份，每次取其中没有测试过的一份作为测试集，依次进行 10 次到所有的链接都被当且仅当只测试过一次。

我们依次改进，CommonNeighbor(CN)链接预测算法为(I_CN1 和 I_CN2)、Jaccard 算法为 I_Jaccard1、Sorensen 算法为 I_Sorensen1、RA 算法为 I_RA 和 Katz 算法，并且使用 10 折交叉验证的方法对传统的链接预测算法和改进的链接预测算法进行了实验，以及分别通过准确率和 AUC 标准度量方法对传统的链接预测算法和改进的链接预测算法进行了度量。

每次实验中，我们对 η 参数的取值都为 0.001、0.01、0.1 和 0.9，对 Katz 方法中的 β 参数的取值范围为 0.00005、0.0005、0.005 和 0.05。

Katz 方法主要分为无权重(UnWeighted Katz)方法和有权重(Weighted Katz)方法，改进的 Katz 方法主要是基于有权重的方法。

以下，是使用 10 折交叉验证方法通过准确率标准度量方法的传统链接预测方法和改进链接预测方法的对比实验结果。

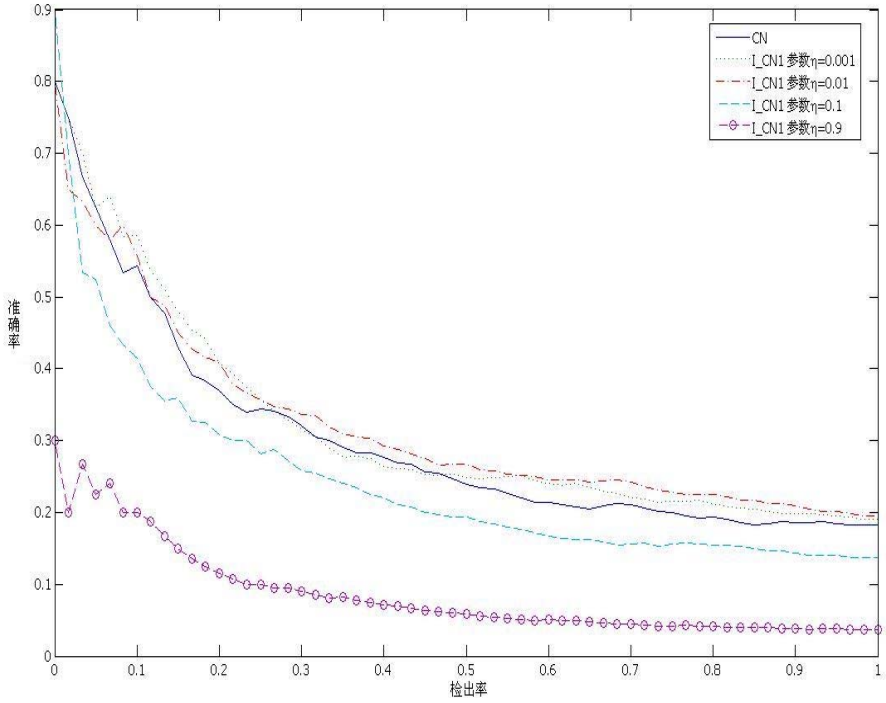

图 5.3 方法 CN 和 I_CN1 的准确率对比图。

Fig 5.3 Precision for the CN and I_CN1 methods.

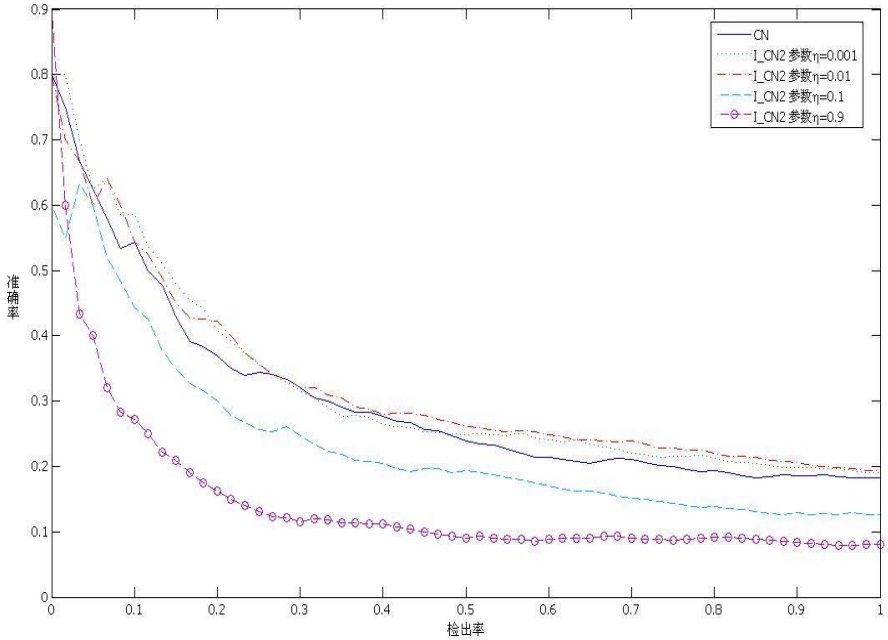

图 5.4 方法 CN 和 I_CN2 的准确率对比图。

Fig 5.4 Precision for the CN and I_CN2 methods.

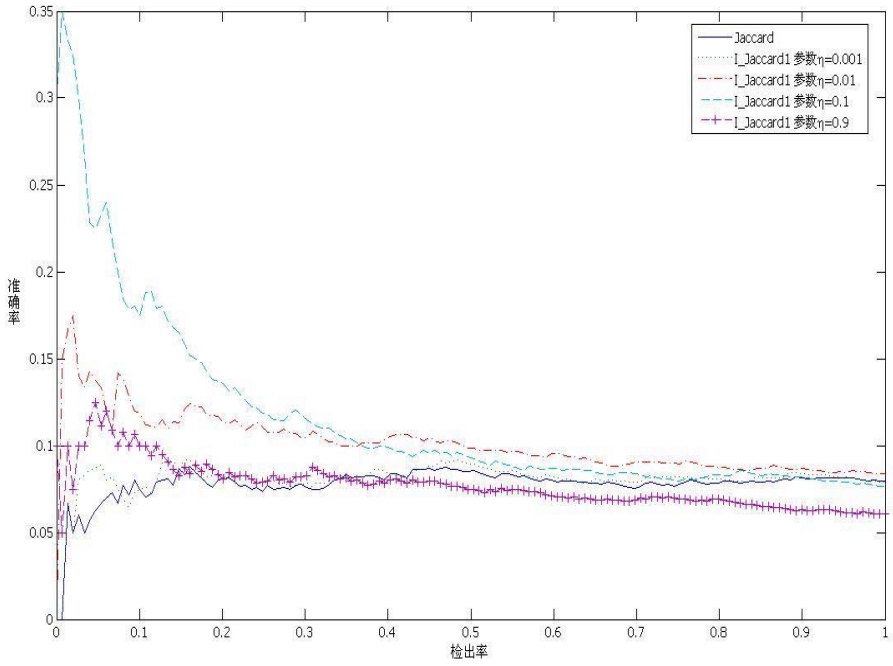

图 5.5 方法 Jaccard 和 I_Jaccard1 的准确率对比图。

Fig 5.5 Precision for the Jaccard and I_ Jaccard1 methods.

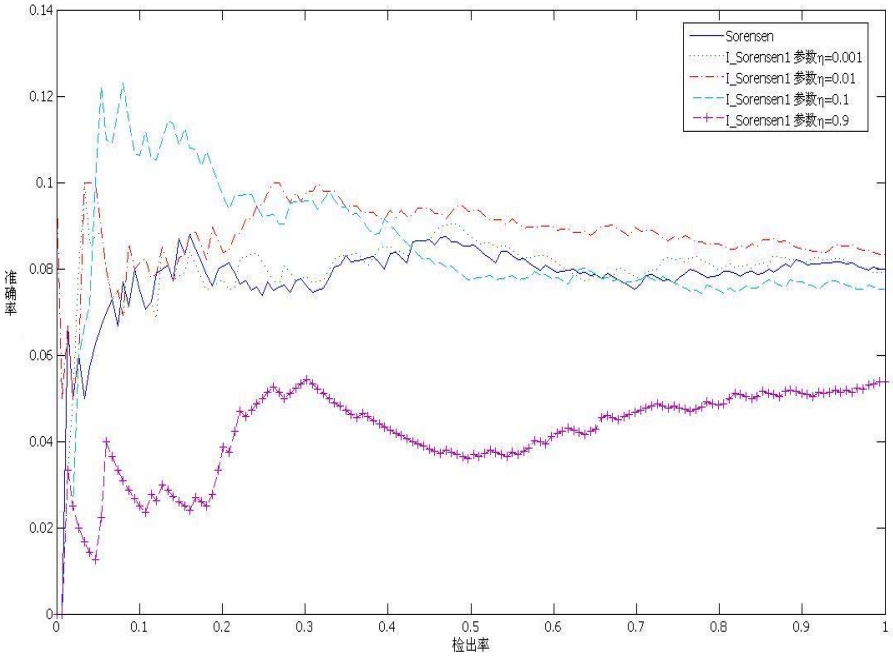

图 5.6 方法 Sorensen 和 I_Sorensen1 的准确率对比图。

Fig 5.6 Precision for the Sorensen and I_ Sorensen1 methods.

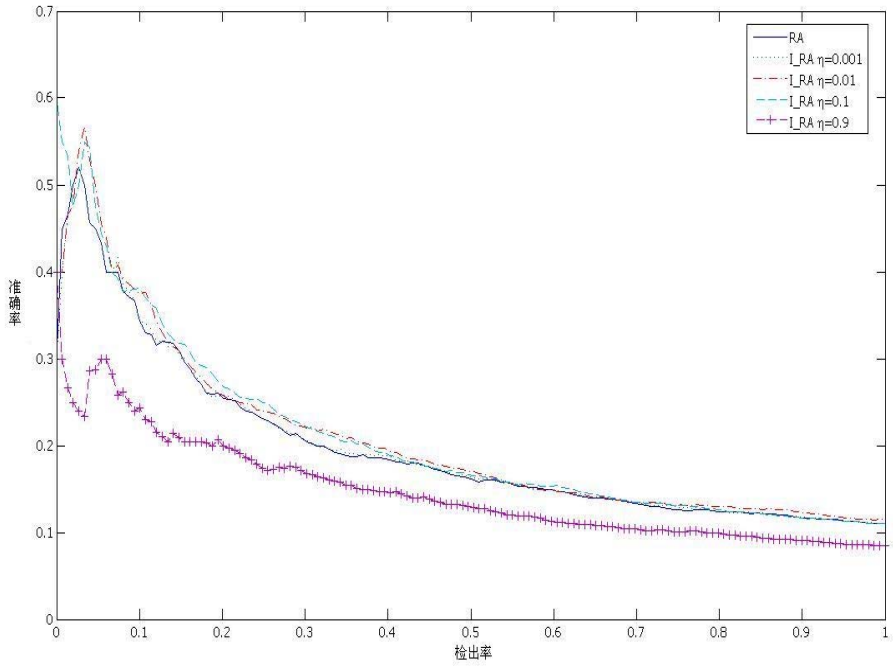

图 5.7 方法 RA 和 I_RA 的准确率对比图。

Fig 5.7 Precision for the RA and I_ RA methods.

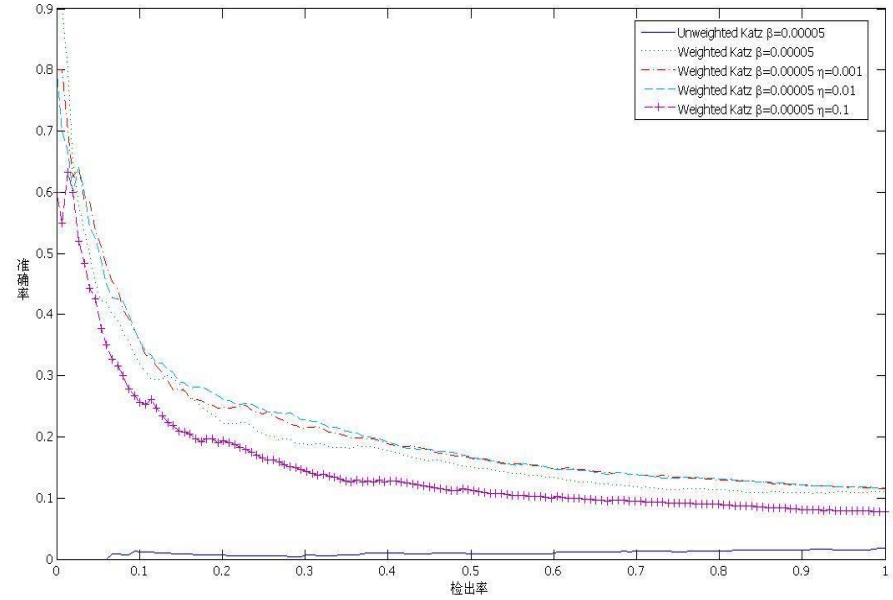

图 5.8 β=0.00005 方法 Katz 和改进 Katz 的准确率对比图。

Fig 5.8 When β is 0.00005, precision for the Katz and Improved Katz methods.

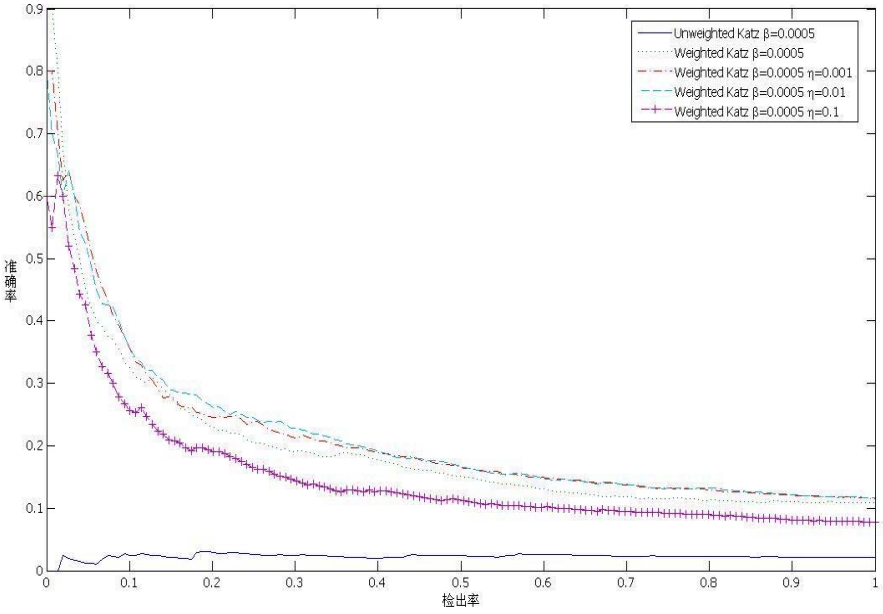

图 5.9 β=0.0005 方法 Katz 和改进 Katz 的准确率对比图。

Fig 5.9 When β is 0.0005, precision for the Katz and Improved Katz methods.

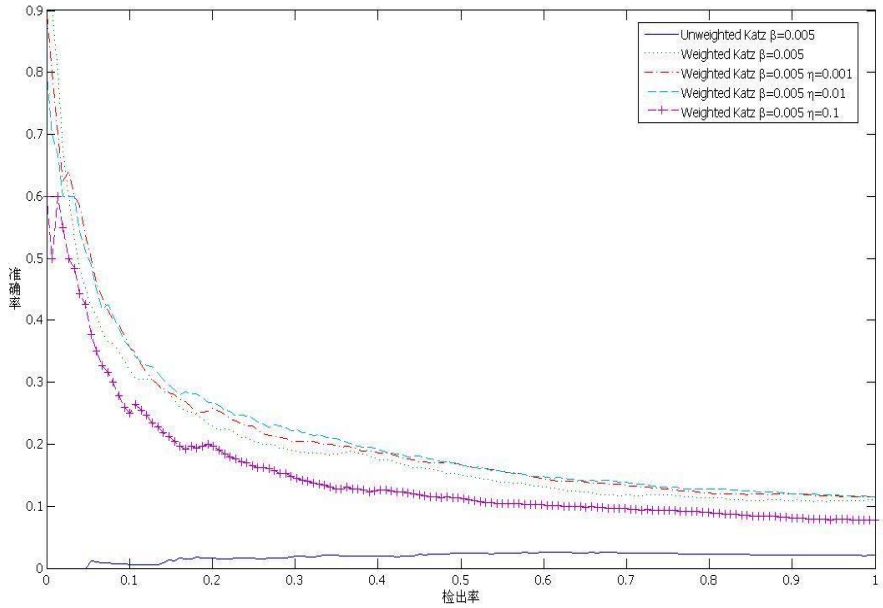

图 5.10 β=0.005 方法 Katz 和改进 Katz 的准确率对比图。

Fig 5.10 When β is 0.005, precision for the Katz and Improved Katz methods.

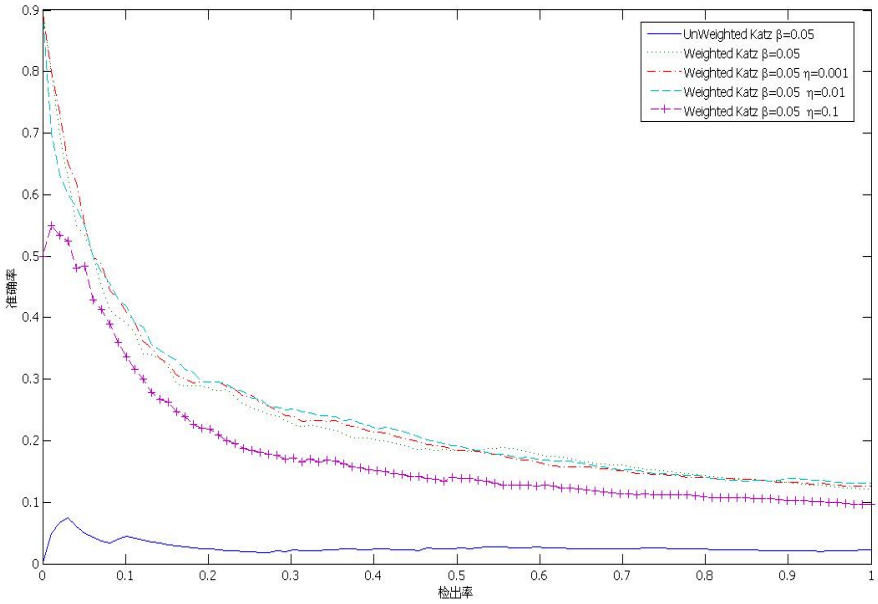

图 5.11 β=0.05 方法 Katz 和改进 Katz 的准确率对比图。

Fig 5.11 When β is 0.05, precision for the Katz and Improved Katz methods.

在表 5.1 中，给出了传统链接预测方法和改进的链接预测方法的 AUC 值。其中，UnWβ 表示在参数 β 取值下，无权重的 Katz 方法，Wβ 表示在参数 β 取值下，有权重的 Katz 方法，Wβη 表示在参数 β、η 取值下，改进的 Katz 链接预测方法。

表 5.1 链接预测方法的 AUC 值。

Table 5.1 The AUC Value for methods of link prediction.

方法	AUC 值	方法	AUC 值
CN	0.652275		
I_CN1 η=0.001	0.695622	I_CN2 η=0.001	0.695914
I_CN1 η=0.01	0.689970	I_CN2 η=0.01	0.697447
I_CN1 η=0.1	0.560490	I_CN2 η=0.1	0.647483
I_CN1 η=0.9	0.463343	I_CN2 η=0.9	0.610589
Jaccard	0.604387		
I_Jaccardη=0.001	0.626104	I_Jaccardη=0.01	0.639879
I_Jaccardη=0.1	0.608283	I_Jaccardη=0.9	0.580353

方法	AUC 值	方法	AUC 值
Sorensen	0.610717		
I_Sorensenη=0.001	0.624739	I_Sorensenη=0.01	0.642248
I_Sorensenη=0.1	0.617872	I_Sorensenη=0.9	0.596165
RA	0.687026		
I_RAη=0.001	0.713309	I_RAη=0.01	0.707960
I_RAη=0.1	0.688012	I_RAη=0.9	0.659176
UnWβ=0.00005	0.754823	Wβ=0.00005	0.889345
Wβ=0.00005η=0.001	0.916064	Wβ=0.00005η=0.01	0.916949
Wβ=0.00005η=0.1	0.901793	Wβ=0.00005η=0.9	0.929893
UnWβ=0.0005	0.800619	Wβ=0.0005	0.874493
Wβ=0.0005η=0.001	0.887930	Wβ=0.0005η=0.01	0.886209
Wβ=0.0005η=0.1	0.880039	Wβ=0.0005η=0.9	0.858471
UnWβ=0.005	0.817258	Wβ=0.005	0.864754
Wβ=0.005η=0.001	0.894879	Wβ=0.005η=0.01	0.883866
Wβ=0.005η=0.1	0.886325	Wβ=0.005η=0.9	0.825696
UnWβ=0.05	0.826676	Wβ=0.05	0.882808
Wβ=0.05η=0.001	0.877223	Wβ=0.05η=0.01	0.877588
Wβ=0.05η=0.1	0.862053	Wβ=0.05η=0.9	0.786718

在图 5.3 和 5.4 中，当动态演化模型参数 η 取值为 0.001 和 0.01 时，改进的 I_CN1 和 I_CN2 共同邻居方法明显优于传统的 CN 共同邻居方法，但是在参数 η 取值为 0.1 和 0.9 时，改进的链接预测方法的效果就明显差于传统的共同邻居方法，同样的在表 5.1 中，改进的 I_CN1 和 I_CN2 链接预测方法的 AUC 度量要明显优于传统链接预测方法。

在图 5.5 中，改进的 I_Jaccard 链接预测方法要明显优于传统的 Jaccard 链接预测方法。在参数 η 取值为 0.1 时，改进的方法最优，但是在参数 η 取值为 0.001 时，改进的链接预测方法逼近传统的链接预测方法。这是由于当参数 η 取值较小时，动态演化模型就近似于时变网络的静态模型，而传统的

Jaccard 链接预测方法都是基于静态模型的。

但是，改进的 I_Jaccard 链接预测方法相比传统的 Jaccard 链接预测方法的 AUC 度量效果却不如准确率的优势那么明显。

从图 5.6 上可以看出，Sorensen 方法的链接预测效果在 Enron 数据集上的效果不明显，尽管 Sorensen 方法的整体效果不好，但是仍可以看出改进的 Sorensen 方法在参数 η 取值为 0.01 和 0.1 时仍明显优于传统链接预测方法。从表 5.1 Sorensen 的 AUC 度量结果也可以看出相似于 Sorensen 方法的准确率结论。

RA 链接预测方法在 Enron 数据集上的链接预测效果很好，从图 5.7 可以看出，RA 方法的准确率曲线明显高于其他链接预测方法。改进的链接预测方法 I_RA 在检出率 0.05 到 0.45 之间的准确率略高于传统的链接预测方法 RA，在其它范围检出率的 RA 方法和改进的 I_RA 方法准确率曲线是逼近状态。

RA 链接预测方法的 AUC 度量值也比较理想，在参数 η 为 0.001 和 0.01 时，改进的链接预测方法 I_RA 略高于传统链接预测方法 RA，参数 η 取值为 0.1 和 0.9 时，改进的 I_RA 链接预测方法与传统链接预测方法 RA 的差别不大。

从表 5.1 中，我们可以明显地看出 Katz 方法在所列的所有方法中，Katz 方法的 AUC 度量值是最高的。并且在参数 β 取值为 0.00005 时，我们可以明显地看到改进的 Katz 链接预测方法要明显高于传统的 Katz 方法。当参数 β 取值为 0.05 时，改进的链接预测方法 Katz 略差于传统的链接预测方法 Katz。

并且从图 5.8，5.9，5.10 和 5.11 可以看出，改进的链接预测方法 Katz 和有权重的 Katz 方法的准确率曲线要明显高于不加权重的传统链接预测方法 Katz，即改进的方法 Katz 和加权重的 Katz 方法在 Enron 数据集上的效果要明显优于不加权重的 Katz 方法，并且改进的 Katz 方法在参数 η 取值为 0.001 和 0.01 时，要明显高于传统带权重的 Katz 方法，而在参数 η 取值为 0.1 和 0.9 时，明显低于传统带权重的 Katz 方法。

从以上的实验分析，我们可以得出一个明显的结论：在参数 η 取合适的值时，改进的链接预测方法要明显优于传统的链接预测方法。此实验也验证了，时变网络的动态演化模型在链接预测中的应用是适合的，并且相比时变网络静态模型用于描述时变网络的动态变化过程要更适合。

此项实验运行在 Windows XP SP3，CPU Core 2 Duo 1.6GHz，内存 1GB 的环境下。传统的链接预测方法，CN，Jaccard，Sorensen，RA 和 Katz 方法，以及改进的链接预测方法，I_CN1，I_CN2，I_Jaccard，I_Sorensen，I_RA 和改进的 Katz 方法，在此平台上的运行时间为 9 分 23 秒钟。

5.4 新的链接预测算法在 Markov 逻辑网中的实验分析

在传统的链接预测方法上，本文提出了基于马尔科夫逻辑网的一个新的链接预测算法 MLP。根据公式(4.11) 和(4.12)，我们主要针对如下几个传统的链接预测方法，CN、Jaccard、Sorensen、Ra 和 Katz 等链接预测方法做了实验验证，并且对每个方法的亲近度离散化为如下几个谓词。

max(f)，表示链接预测方法 f 的最大亲近度；min(f)，表示链接预测方法 f 的最小亲近度。$length = \max(f) - min(f)$，$length$表示最大亲近度和最小亲近度之间的差值。

传统的 CN 方法应用于马尔科夫逻辑网，主要离散化为谓词 Cn0、Cn1 和 Cn2。$length = \max(Cn) - min(Cn)$。

谓词 Cn0(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Cn) \le Score(x,y) < \min(Cn) + 0.33 \times length$，否则为假。

谓词 Cn1(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Cn) + 0.33 \times length \le Score(x,y) < \min(Cn) + 0.66 \times length$，否则为假。

谓词 Cn2(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Cn) + 0.66 \times length \le Score(x,y) \le \max(Cn)$，否则为假。

传统 Jaccard 方法应用于马尔科夫逻辑网，主要离散化为谓词 Jaccard0 、Jaccard1、Jaccard2 和 Jaccard3。$length = \max(\text{Jaccard}) - min(\text{Jaccard})$。

谓词 Jaccard0(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Jaccard) \le Score(x,y) < \min(Jaccard) + 0.25 \times length$，否则为假。

谓词 Jaccard1(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Jaccard) + 0.25 \times length \le Score(x,y) < \min(Jaccard) + 0.5 \times length$，否则为假。

谓词 Jaccard2(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Jaccard) + 0.5 \times length \le Score(x,y) < \min(Jaccard) + 0.75 \times length$，否则为假。

谓词 Jaccard3(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Jaccard) + 0.75 \times length \le Score(x,y) \le \max(Jaccard)$，否则为假。

传统的 Sorensen 方法应用于马尔科夫逻辑网，主要离散化为 4 个谓词 Sorensen0 、Sorensen1、Sorensen2 和 Sorensen3。其中亲近度的差值表示，如式$length = \max(\text{Sorensen}) - min(\text{Sorensen})$。

谓词 Sorensen0(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Sorensen) \le Score(x,y) < \min(Sorensen) + 0.25 \times length$，否则为假。

谓词 Sorensen1(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Sorensen) + 0.25 \times length \le Score(x,y) < \min(Sorensen) + 0.5 \times length$，否则为假。

谓词 Sorensen2(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Sorensen) + 0.5 \times length \le Score(x,y) < \min(Sorensen) + 0.75 \times length$，否则为假。

谓词 Sorensen3(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(Sorensen) + 0.75 \times length \le Score(x,y) \le \max(Sorensen)$，否则为假。

传统 RA 方法应用于马尔科夫逻辑网，主要离散化为谓词 RA0 、RA1、RA2 和 RA3。$length = \max(RA) - min(RA)$。

谓词 RA0(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(RA) \le Score(x,y) < \min(RA) + 0.25 \times length$，否则为假。

谓词 RA1(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(RA) + 0.25 \times length \le Score(x,y) < \min(RA) + 0.5 \times length$，否则为假。

谓词 RA2(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(RA) + 0.5 \times length \le Score(x,y) < \min(RA) + 0.75 \times length$，否则为假。

谓词 RA3(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(RA) + 0.75 \times length \le Score(x,y) \le \max(RA)$，否则为假。

传统 Katz 方法应用于马尔科夫逻辑网，无权重的方法离散化为如下谓词，UnWKatzij(x, y)，其中$\{0 \le j \le 5 | j \in N\}$，$i = 0$，表示参数 β=0.00005 时，$i = 1$，表示参数 β=0.0005 时，$i = 2$，表示参数 β=0.005 时，$i = 3$，表示参数 β=0.05 时，分别各自离散化为 6 个不同的谓词，同理，WKatzij(x, y)，其中$i = 0$，表示参数 β=0.00005 时，$i = 1$，表示参数 β=0.0005 时，$i = 2$，表示参数 β=0.005 时，$i = 3$，表示参数 β=0.05 时，也分别各自离散化为 6 个不同的谓词。

$length = \max(kz) - min(kz)$，对于谓词 UnWKatzij(x, y)，WKatzij(x, y)的解释，我们用符号 kzj 分别代表参数 β 某值下的无权重或有权重 Katz 方法的第 j 个谓词。

谓词 kzj(x, y)为真，当且仅当对象x和y的亲近度，满足 $\min(kz) + (j/6) \times length \le Score(x,y) < \min(kz) + ((j+1)/6) \times length$，否则为假。

在表 5.2 中，给出了我们建立的马尔科夫逻辑网，并通过 Enron 数据集学习到了每个逻辑子句的权重。

表 5.2 新链接预测方法的马尔科夫逻辑网。

Table 5.2 The MLN of the new method for link prediction.

逻辑子句	权重	备注
Cn0(x,y) => Link(x,y)	**1.279070**	
Cn1(x,y) => Link(x,y)	0.064836	

逻辑子句	权重	备注
Cn2(x,y) => Link(x,y)	0.304003	
Jaccard0(x,y) => Link(x,y)	-0.590957	
Jaccard1(x,y) => Link(x,y)	0.683911	
Jaccard2(x,y) => Link(x,y)	0.030189	
Jaccard3(x,y) => Link(x,y)	-0.182790	
Sorensen0(x,y) => Link(x,y)	0.119439	
Sorensen1(x,y) => Link(x,y)	-0.207407	
Sorensen2(x,y) => Link(x,y)	0.134912	
Sorensen3(x,y) => Link(x,y)	0.330719	
RA0(x,y) => Link(x,y)	-0.426051	
RA1(x,y) => Link(x,y)	0.417617	
RA2(x,y) => Link(x,y)	0.253256	
RA3(x,y) => Link(x,y)	0.168028	
UnWKatz00(x,y) => Link(x,y)	**2.327890**	**参数 β=0.00005**
UnWKatz01(x,y) => Link(x,y)	-0.829832	参数 β=0.00005
UnWKatz02(x,y) => Link(x,y)	0.590298	参数 β=0.00005
UnWKatz03(x,y) => Link(x,y)	0.482334	参数 β=0.00005
UnWKatz04(x,y) => Link(x,y)	-0.333813	参数 β=0.00005
UnWKatz05(x,y) => Link(x,y)	0.072204	参数 β=0.00005
UnWKatz10(x,y) => Link(x,y)	0.419990	参数 β=0.0005
UnWKatz11(x,y) => Link(x,y)	-0.628297	参数 β=0.0005
UnWKatz12(x,y) => Link(x,y)	0.662790	参数 β=0.0005
UnWKatz13(x,y) => Link(x,y)	0.586048	参数 β=0.0005
UnWKatz14(x,y) => Link(x,y)	0.211860	参数 β=0.0005
UnWKatz15(x,y) => Link(x,y)	0.072204	参数 β=0.0005
UnWKatz20(x,y) => Link(x,y)	-0.377419	参数 β=0.005
UnWKatz21(x,y) => Link(x,y)	*-1.702220*	*参数 β=0.005*
UnWKatz22(x,y) => Link(x,y)	0.933163	参数 β=0.005
UnWKatz23(x,y) => Link(x,y)	1.066600	参数 β=0.005
UnWKatz24(x,y) => Link(x,y)	0.561116	参数 β=0.005
UnWKatz25(x,y) => Link(x,y)	0.072204	参数 β=0.005
UnWKatz30(x,y) => Link(x,y)	*-1.732710*	*参数 β=0.05*
UnWKatz31(x,y) => Link(x,y)	*-1.120840*	*参数 β=0.05*

逻辑子句	权重	备注
UnWKatz32(x,y) => Link(x,y)	**1.533040**	**参数 β=0.05**
UnWKatz33(x,y) => Link(x,y)	**1.489640**	**参数 β=0.05**
UnWKatz34(x,y) => Link(x,y)	**1.516690**	**参数 β=0.05**
UnWKatz35(x,y) => Link(x,y)	0.072204	参数 β=0.05
WKatz00(x,y) => Link(x,y)	0.132466	参数 β=0.00005
WKatz01(x,y) => Link(x,y)	0.064836	参数 β=0.00005
WKatz02(x,y) => Link(x,y)	0.280165	参数 β=0.00005
WKatz03(x,y) => Link(x,y)	0.355196	参数 β=0.00005
WKatz04(x,y) => Link(x,y)	0.800788	参数 β=0.00005
WKatz05(x,y) => Link(x,y)	**1.295680**	**参数 β=0.00005**
WKatz10(x,y) => Link(x,y)	-0.473682	参数 β=0.0005
WKatz11(x,y) => Link(x,y)	0.064836	参数 β=0.0005
WKatz12(x,y) => Link(x,y)	0.280165	参数 β=0.0005
WKatz13(x,y) => Link(x,y)	0.355196	参数 β=0.0005
WKatz14(x,y) => Link(x,y)	0.800788	参数 β=0.0005
WKatz15(x,y) => Link(x,y)	**1.295680**	**参数 β=0.0005**
WKatz20(x,y) => Link(x,y)	-0.775056	参数 β=0.005
WKatz21(x,y) => Link(x,y)	0.064836	参数 β=0.005
WKatz22(x,y) => Link(x,y)	0.280165	参数 β=0.005
WKatz23(x,y) => Link(x,y)	0.355196	参数 β=0.005
WKatz24(x,y) => Link(x,y)	0.800788	参数 β=0.005
WKatz25(x,y) => Link(x,y)	**1.295680**	**参数 β=0.005**
WKatz30(x,y) => Link(x,y)	*-1.497040*	*参数 β=0.05*
WKatz31(x,y) => Link(x,y)	-0.500288	参数 β=0.05
WKatz32(x,y) => Link(x,y)	0.146849	参数 β=0.05
WKatz33(x,y) => Link(x,y)	0.048638	参数 β=0.05
WKatz34(x,y) => Link(x,y)	0.743137	参数 β=0.05
WKatz35(x,y) => Link(x,y)	0.045163	参数 β=0.05

从表 5.2 中可以看出，马尔科夫逻辑网中加粗显示的逻辑子句属于偏硬规则。如果满足了加粗显示的这些逻辑子句，则两个对象之间发生链接的可能性就很大，而倾斜显示的这些逻辑子句被满足，则两个对象之间发生链接的可能性很小。

同样，我们也是将 Enron 实验数据集随机分成 10 份，每次取其中没有测试过的一份作为测试集，依次进行 10 次到所有的链接都被当且仅当只测试过一次。

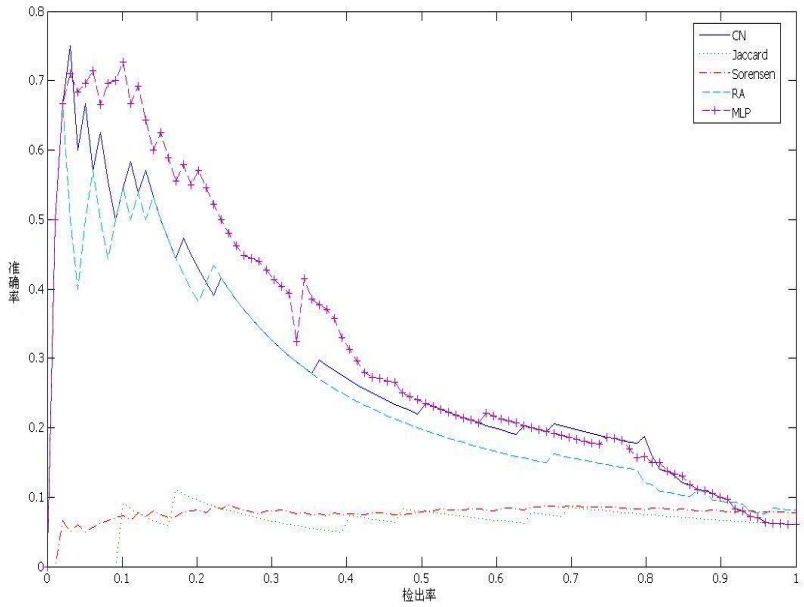

图 5.12 方法 CN、Jaccard、Sorensen、RA 和 MLP 的准确率对比图。

Fig 5.12 Precision for the CN, Jaccard, Sorensen, RA and MLP methods.

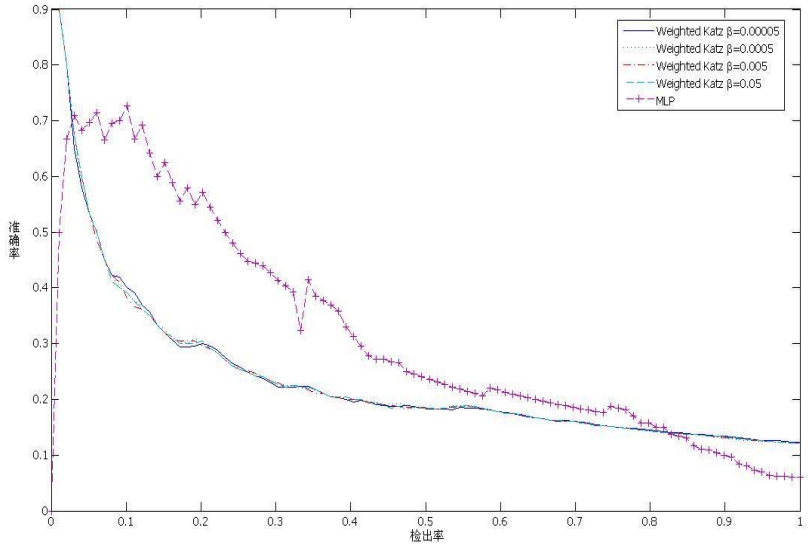

图 5.13 加权重的方法 Katz 和 MLP 的准确率对比图。

Fig 5.13 Precision for the weighted Katz and MLP methods.

在每次的数据集预处理中，我们给出所有 CN、Jaccard、Sorensen、RA 和 Katz 方法构成谓词的所有解释，以及训练集组成的实例库，利用 Alchemy 中的 Infer 工具，计算出所有未发生链接的边的概率。我们根据未发生链接边的概率进行降序排列，然后进行链接预测。

在实验方法中，我们采用 10 折交叉验证的方法，利用准确率的标准度量方法对提出的新的链接预测方法进行分析。图 5.12 和图 5.13，给出了提出的新的链接预测方法的准确率度量标准图。图中 MLP 是在本文中提出的一种新的链接预测算法。

从图 5.12 中，我们可以看到新的链接预测算法 MLP 的准确率，在检出率 0.05 到 0.5 之间均高于 CN、Jaccard、Sorensen 和 Katz 等传统的链接预测算法，在其他的检出率下，略低于传统链接预测算法 CN 和 RA，在基本上所有的检出率下，均高于传统链接预测算法 Jaccard 和 Sorensen。从图 5.13 中，我们可以看到 MLP 链接预测算法的准确率，在 0.1 到 0.8 之间明显高于不同参数的加权 Katz 链接预测算法，在其他检测率下，略低于加权的 Katz 链接预测算法。由于从图 5.8、5.9、5.10 和 5.11 中，不加权的链接预测算法 Katz 的准确率明显低于加权的链接预测算法，所以不与 MLP 链接预测算法进行比较。

从上面的实验分析，我们可以得到一个明显的结论：提出的新的链接预测算法 MLP 明显要优于传统的链接预测算法。

该实验是在 pentium4 处理器，主频 3.0GHz，内存 2GB，操作系统 Red Hat Enterprise Linux 6 的环境下进行的，其中，马尔科夫逻辑网权重的学习过程时间为 2 小时 24 秒，马尔科夫逻辑网的推理过程时间为 7 分钟 19 秒。

5.5 本章小结

本章主要对基于时变网络的动态演化模型的改进的链接预测方法和基于马尔科夫逻辑网的新的链接预测方法，做了对比实验和进行了实验分析。我们对本文改进的和新的链接预测方法，做了 10 折交叉验证的实验，得到的实验结果客观地反映了算法的性能。从实验的结果和分析中，我们可知改进的链接预测算法和新的链接预测算法明显优于传统链接预测方法。

6 总结与展望

6.1 总结

链接预测研究是一个最新广泛研究的课题。研究学者们提出了许多不同的链接预测方法，主要分为基于网络结构和概率关系模型的两类方法。在本文中，根据以前基于网络结构的研究提出了一个动态演化模型来精确量化对象之间关系的动态变化过程，并且在此模型上改进了链接预测方法，通过实验对比，改进的链接预测算法要明显优于传统的链接预测算法。同时，我们将链接预测的研究结合最新的概率关系模型马尔科夫逻辑网 MLN 进行新的研究，并在此模型上提出了新的链接预测算法，以及通过实验分析，新的链接预测算法要明显优于传统的链接预测算法。

总结本文，我的主要研究内容如下。

① 由于经典链接预测算法都是基于时变网络的静态模型，本文提出一种对社交网络等时变网络随时间演化的动态演化模型。

② 本文对一些传统链接预测算法进行了改进，使适合于时变网络的动态演化模型，经过改进的链接预测算法对时变网络的链接预测准确率有明显地提高。

③ 我将对不同性质的数据集各有不同优缺点的传统链接预测算法，融入马尔科夫逻辑网模型，并且基于该模型提出新的链接预测算法。

6.2 展望

虽然，本文中基于动态演化模型改进的链接预测算法和基于马尔科夫逻辑网提出的新的链接预测方法都优于传统的链接预测算法，但是，其中还有许多亟待研究的问题还没有解决，对于进一步提升本文中创新链接预测算法的准确率形成了新的瓶颈。

在时变网络的动态演化模型中，如何优化参数 η 是一个棘手的问题。因为 η 的大小，对于模型描述时变网络的动态变化过程影响极大。参数 η 取值的研究，是将来提高时变网络动态演化模型的描述能力的一个主要研究问题。

在新的链接预测算法中，如何量化传统链接预测方法的亲近度，会严重影响到准确率。另外，如果对于传统链接预测方法的亲近度量化过细，则会严重影响新的链接预测算法的执行效率。所以对于如何解决传统链接预测方法亲近度的量化问题的研究，将会是一个亟待解决的研究问题。

致　　谢

在三年的学习生活中，我的老师、同学和家人给了我最有力的支持和鼓励，在此我对他们表示最真挚的感谢！

在我的导师邢永康副教授的精心指导和悉心关怀下，本文的主要研究工作得以顺利完成。在我的学习素养培养和论文的研究工作中，无不映衬着导师辛勤教育、指导的身影。导师的严谨治学态度、渊博的知识、生动形象的传授，使我深受启迪。从尊敬的导师身上，我不仅学到了扎实、宽广的专业知识，也学到了做学问的方法和做人的道理。在此我要向对我求学路上影响深远的导师，致以最衷心的感谢和深深的敬意。

在这短短三年的学习生活中，有挫折，也有欢笑，非常感谢给予我帮助的朋友们，特别感谢我实验室的兄弟姐妹，刘小军、袁文群和项卫平，在论文创作中，给予了很多帮助。

在此，向所有关心和帮助过我的领导、老师、同学和朋友表示衷心地感谢！

最后，衷心地感谢各位专家、教授，在百忙之中抽出时间评阅我的论文和参加我的答辩！

武南南

二O一二年四月　于重庆

参考文献

[1] L Getoor, C P Diehl. Link mining: a survey [J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.

[2] L Getoor. Link mining: a new data mining challenge [J]. SIGKDD Explorations, 2003, 5(1):84-89.

[3] J O'Madadhain, J Hutchins, P Smyth. Prediction and ranking algorithms for even-based network data [J]. SIGKDD Explorations, 2005, 7(2).

[4] M Rattigan, D Jensen. The case for anomalous link discovery [J]. SIGKDD Explorations, 2005, 7(2).

[5] D Liben-Nowell, J Kleinberg. The link prediction problem for social networks [C]. In International Conference on Information and Knowledge Management(CIKM), 2003: 556-559.

[6] H Kautz, B Selman, M Shah. ReferralWeb: Combining social networks and collaborative filtering [J]. Communications of the ACM, 1997, 30(3).

[7] P Raghavan. Social networks: From the web to the enterprise [J]. IEEE Internet Computing, 2002, 6(1):91-94.

[8] Valdis Krebs. Mapping networks of terrorist cells [J]. Connections, 2002, 24(3):43–52.

[9] Glen Jeh, Jennifer Widom. SimRank: A measure of structural-context similarity [C]. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, NY, USA, 2002:538-543.

[10] Leo Katz. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1):39-43.

[11] MEJ Newman. Clustering and preferential attachment in growing networks [J]. Physical Review Letters E, 2001, 64(025102).

[12] Nesserine Benchettara, Rushed Kanawati, et al. Supervised Machine Learning applied to Link Prediction in Bipartite Social Networks [C]. International Conference on Advances in Social Networks Analysis and Mining, Villetaneuse France, 2010: 326-330.

[13] M Craven, D DiPasquo, D Freitag, et al. Learning to construct knowledge bases from the world wide web [J]. Artificial Intelligence, 2000, 118(1-2):69-114.

[14] J O'Madadhain, P Smyth. EventRank: A framework for ranking time-varying networks [C]. In Proceedings of the 3rd international workshop on Link discovery, NY USA, 2005:9-16.

[15] J O'Madadhain, P Smyth, L Adamic. Learning predictive models for link formation [C].

Presented at the International Sunbelt Social Network Conference, CA, USA, 2005.

[16] A Popescul, L H Ungar. Statistical relational learning for link prediction [C]. In IJCAI Workshop on Learning Statistical Models from Relational Data, Acapulco, Mexico, 2003.

[17] B Taskar, MF Wong, P Abbeel, et al. Link prediction in relational data [C]. In Proceedings of Neural Information Processing Systems (NIPS), Vancouver, Canada, 2003.

[18] R Chellappa, A Jain. Markov random fields: theory and applications [J]. Academic Press, Boston, 1993.

[19] L Getoor, N Friedman, D Koller, et al. Learning probabilistic models of link structure [J]. Journal of Machine Learning Research, 2002, 3:679-707.

[20] Wu Junying, Xia Chunhe, Lv Liangshuang, et al. GLP a Group Link Prediction Algorithm in DTMNs [C]. International Corerence on Educational and Information Technolog (T 2010), Chongqing, China, 2010:V2197-V2201.

[21] CD Manning, P Raghavan, H Schutze. Introduction to Information Retrieval [M]. Cambridge University Press, New York, 2008.

[22] Vartak S. A survey on link prediction [R]. http://sourabhvartak.com/pdf/A Survey on Link Prediction.pdf, 2008.

[23] MEJ Newman. Clustering and preferential attachment in growing networks [J]. Physical Review Letters E, 2001, 64(025102): 251021-251024.

[24] S Carmi, S Havlin, S Kirkpatrick, et al. A model of Internet topology using k-shell decomposition [C]. Proc, Natl. Acad. Sci. USA, 2007.

[25] T Murata, S Moriyasu. Link Prediction of Social Networks Based on Weighted Proximity Measures [C]. IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, CA, United states, 2007: 85-88.

[26] T Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [J]. Biol. Skr. 5, 1948.

[27] EA Leicht, P Holme, MEJ Newman. Vertex similarity in networks [J]. American Physical Society, 2006, 73(2).

[28] G Salton, MJ McGill. Introduction to Modern Information Retrieval [M]. McGraw-Hill, 1983.

[29] LA Adamic, E Adar. Friends and neighbors on the web [J]. Social Networks, 2003, 25(3):211-230.

[30] S Brin, L Page. The anatomy of a large-scale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7):107-117.

[31] Linyuan Lü, Tao Zhou. Role of weak ties in link prediction of complex networks [C]. Proceeding of the 1st ACM international workshop on Complex networks meet information & knowledge management, Hong Kong, China, 2009:55-58.

[32] Linyuan Lü, Tao Zhou. Link prediction in complex networks [J]: A survey. Physica A, 2011, 390(2011):1150-1170.

[33] A Clauset, C Moore, MEJ. Newman. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453:98-101.

[34] N Friedman, L Getoor, D Koller, et al, Learning probabilistic relational models [C]. in: Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999.

[35] B Taskar, P Abbeel, D Koller. Discriminative probabilisticmodels in relational data [C]. in: Preceedings of the 18th Conference onUncertainty in Artificial Intelligence, UAI02, Edmonton, Canada, 2002:485-492.

[36] P Domingos, M Richardson. Markov logic: A unifying framework for statistical relational learning [C]. In ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields, 2004:49-54.

[37] RN Lichtenwalter, JT Lussier, NV Chawla. New perspectives and methods in link prediction [C]. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, 2010:25-28.

[38] MEJ Newman. The structure of scientific collaboration networks [J]. Proceedings of the National Academy of Sciences, 2001, 98:404-409.

[39] Evrim Acar, DM. Dunlavy, TG. Kolda. Link Predictionon Evolving Data using Matrix and Tensor Factorizations [C]. IEEE International Conference on Data Mining Workshops, Miami, FL, United States, 2009:262-269.

[40] RJ Crane, Luke K McDowell. Evaluating Markov Logic Networks for Collective Classification [D]. National Technical Information Service, Naval Academy, Annapolis, MD. Dept. of Computer Science. 2010.

[41] RJ Crane, LK. McDowell. INVESTIGATING MARKOV LOGIC NETWORKS FOR COLLECTIVE CLASSIFICATION [C]. Proceedings of the Fourth International Conference on Agents and Artificial Intelligence (ICAART 2012), 2012.

[42] J Shetty, J Adibi. The Enron email dataset database schema and brief statistical report [R]. Information Sciences Institute Technical Report, University of Southern California. 2004.

[43] B Klimt , Y Yang. The enron corpus: A new dataset for email classification research [C]. Proceedings of 15th European Conference on Machine Learning, Pisa, Italy, 2004: 217-226.

[44] M Sumner, P Domingos. The Alchemy Tutorial. http://alchemy.cs.washington.edu. 2010-3-20.

[45] M Bilgic, G M Namata, L Getoor. Combining Collective Classification and Link Prediction [C]. Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, University of Maryland, USA, 2007: 381-386.

[46] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]. Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publisher, Quebec, Canada, 1995: 1137-1143.

附　　录

A. 作者在攻读硕士学位期间发表的论文目录：

[1] Xing Yongkang, Wu Nannan. Dynamic Ranking Objects based on Time Varying Event Network. JICSIT 2011. (EI 检索，已录用).