南京邮电大学

毕业设计(论文)外文资料翻译

|  |  |
| --- | --- |
| 学 院 | 计算机学院、软件学院 |
| 专　　业 | 软件工程（NIIT） |
| 学生姓名 | 陈志聪 |
| 班级学号 | B12041120 |
| 外文出处 |  |

附件：1.外文资料翻译译文；2.外文原文

|  |
| --- |
| 指导教师评价：  1．翻译内容与课题的结合度： □ 优 □ 良 □ 中 □ 差  2．翻译内容的准确、流畅： □ 优 □ 良 □ 中 □ 差  3．专业词汇翻译的准确性： □ 优 □ 良 □ 中 □ 差  4．翻译字符数是否符合规定要求：□ 符合 □ 不符合  　　　　　　　　　　　　　　　　　　　　　　　指导教师签名：  　　　　　　　　　　　　　　　　　　　　　　　　　　 年　　月　　日 |

附件1：外文资料翻译译文

社会关系的离散度与情侣关系——对Facebook上人际关系状态的网络分析

摘要（Abstract）

在分析在线社会网络时，一个重要的工作就是识别那些在单个个体的邻居网络中有着强社交关系的关键人物。在本文中，我们从情侣这一强关系入手研究这个问题。我们的研究仅仅围绕着这个问题展开：给定一个人与他所有的朋友之间的关系网络，是否可以仅从网络的结构角度，从这些朋友中识别出谁和他是情侣关系？利用Facebook用户产生的大量样本数据，我们发现这个想法可以实现，且识别精度较高。这其中的关键在于，我们引入了一个新的概念：离散度（dispersion），它表征了两个人的共同好友之间缺乏相连关系的程度。实验结果为识别在线应用中那些结构重要性高的人群提供了方法，也给出了现有社交关系强度理论的一个拓展的思路。

简介（Introduction）

在社会网络中，由某个个体及与其相连的邻居节点构成的邻居网络中包含的信息在很多领域的应用中得以展现，例如社会支持（social support）和就业机会（professional opportunities）。现在，人们越来越依赖于社交网络去经营其生活的方方面面，因此他们的线上社交网所包含的信息越来越多，复杂性也随之增强。

宏观上来说，一个人社交网络中的邻居包含了多种多样的关系，包括家庭关系、同事关系、长期的朋友关系，一面之交，情侣关系等等。因此分析这样的社会网络关系的关键就在于利用现有的数据去认知这丰富多样的社交关系。如果掌握了这样一套行之有效的方法，可以在处理线上人际关系上有诸多应用，如个性化的内容推荐，提高线上社交关系的凝聚力等等。

关系强度（Tie Strength）

关系强度是描述一个人与其他人的社交关系的一个重要维度。关系强度非正式地反映了一个人与他人的亲密程度。它上至与亲密朋友之间的强关系，下到与更多一面之交的弱关系。之前在社会学领域已经有对关系强度和他们在社会网络中的结构属性的研究。强关系通常“嵌入”在网络中，与之相关的大量节点之间都是朋友关系，并且经常同时出现大量分享等互动行为。相反的，弱关系通常出现在少量的朋友关系之间，它往往在网络中两个不同部分之间搭起“桥梁”。

与理解强关系相连的一个基本问题就是，通过一个人的社交网络中的潜在结构来认识他的社交网络中最重要的那些人。那么在一个人的强关系中如何定义其结构特征，我们又该如何认识它们呢？解决好这个问题，将大大地帮助个人组织好其线上关系，同时帮助我们以更广阔的视角观察亲密关系的社交网络中的结构。

近来，基于E-mail、手机通话以及社交媒体等数据，已经有一些研究开发了一些分析和估计在线社交关系强度的方法。在这些研究中，最核心的结构性特征是“嵌入度”（embeddedness），它是两个人之间共同好友的数量，通常随着关系强度的增长而增长。事实上，嵌入度和关系强度如此紧密的联系在一起，这使得我们反思，除了嵌入度，是否还有其他的结构度量，能够更准确地描述特定的强关系。

浪漫关系（Romantic Relationships）

我们的工作为亲密关系（如情侣和配偶关系）提出了一种新的基于网络的特征描述。我们研究这样一种亲密关系基于如下的考虑。

从一个现实的角度看，在一个人从青春期到老年的一生的社交过程中，浪漫关系是一种仅有的非一般的强社交关系。它对于一个人日常的行为以及在社交媒体上的表现均有重要的影响。

从方法论的角度来说，这种浪漫关系几乎是一种最强的强关系，但对于这样一种关系，依然没有清楚的结构理论能够基于“嵌入度”对它进行充分的描述。亦或者它们是否本身就拥有属于浪漫关系本身的独特的结构属性。

我们的核心发现在于，嵌入度确实是一种较弱的描述浪漫关系的办法，而我们所提出的离散度（dispersion）相比之下则更加有效。我们对离散度的测量不仅仅基于两个人共同好友的数量，更关注两个人共同好友之间的网络结构，即如果两个人的共同好友之间的连接程度较低，则他们俩之间具有较高的离散度。

基于大量Facebook用户数据（在用户信息中标记着是否有情侣关系）构建的随机样本，我们发现，在识别浪漫关系的问题上，与使用嵌入度相比，采用离散度作为衡量标准，可以获得两倍的精确度。事实上，对于已婚的Facebook用户,我们这种基于网络结构的测算方法，比一个使用机器学习方法，加入诸多用户交互的数据记录作为属性（包括消息、评论、页面浏览量、在事件和照片中同时出现的频率等）构建的复杂分类器的分类效果要好。更进一步地，通过将离散度和这些用户交互特征相结合，可以取得更高的识别精度。

本文的主要贡献如下：

1. 我们提出一个新的网络测度，离散度，用来估算关系强度。现有的对关系强度的分析大多依赖于嵌入度，而离散度的引入给出了一种新方法和思路，用于测算关系强度及其在线上领域的分类问题。
2. 基于我们对浪漫关系对用户在社交网络中行为的影响的理解，我们提供一个新的实质性的基于网络结构来描述浪漫关系的方法。
3. 基于这个特性，我们变化不同的条件和数量进行灵敏度分析。例如，我们发现，通过网络结构识别出来的结果有显著的性别差异，此外，通过这种方式识别出来的情感关系更稳定更坚固。

同样重要的是要描绘我们结果的范围。我们的方法分析在线社会浪漫的伙伴关系，主要是通过它们对网络结构的影响，也因此可以从网络结构特征反过来认识这种浪漫关系。因此，一种潜在的办法是结合其他关于这类关系在线上如何表达的观点，完整地描绘出人们在社会化媒体上的活动规律。

数据与问题描述（Data and problem description）

我们分析社交网络上的浪漫关系所使用的数据集为从Facebook上标明其与他人情侣关系的用户集中随机选择，这包含了已婚（married）、订婚（engaged）、恋爱中（in a relationship）几种。为了从最基础开始验证不同的结构性理论，我们从一个最简单的预测问题开始思考：我们选取了某一个Facebook用户，他已经声明了他有个伴侣，现在我们隐去这个伴侣是谁，仅给出与他相连的关系网络——他的所有朋友们及他们互相之间的连接关系，仅通过这样的网络结构信息，我们能够多么精确的识别出来谁是他的伴侣？图1给出了这样一个Facebook用户的关系网络示例，该用户在网络中心，这样的网络将作为我们模型的输入，我们希望从这样一个网络中预测出谁是他的另一半。也即我们通过他实际填写的信息来验证通过网络结构信息识别出来的情侣关系是否准确。

我们注意到我们的问题有一个重要的特点：鉴于一个用户已经宣布和另一个人的情侣关系，我们想要知道我们可以多么有效地发现这个伙伴。与之相关的还有一些其他的问题，例如，从用户的网络社交关系推断他是否正在恋爱中。在后续部分我们会简要讨论这些问题之间的联系，但如何确定的这样的伙伴关系是我们主要关注的重点。

对于我们分析所用的数据，我们主要使用来自于Facebook网络社区的两个集合。第一个是大的网络社区，从大约130万的Facebook的用户中进行均一的随机选择，所有用户的年龄至少20，拥有约50到2000个朋友，他们均在用户信息中标明了他们的配偶或伴侣关件列表系。由每个用户构成的小的ego network平均有291个节点和6652个链接，对于整体数据集包含大约3.79亿个节点和86亿的链接。

我们也使用一个较小的数据集——从第一个数据集中选出73000个ego网络，进行均匀随机抽样，每个ego网络包含最多25000链接。我们将这个示例为主数据集（primary dataset），之前那个更大的数据集称为扩展数据集（extended dataset）。我们对结构性特征和用户交互特征的计算均基于这两个数据集，这些测算两个数据集上表现出几乎相同的性能。正如我们下面进一步讨论的那样，我们也评估了额外的网络度量，以及基于机器学习算法的更复杂的组合的措施，但那只基于主数据集。

所有这些分析使用匿名Facebook的数据，且是总体上完成的。

为了评价我们从网络结构角度识别浪漫关系的方法，我们从最基本的开始，即从嵌入度研究关系强度。嵌入度指网络中两个人的共同好友个数。嵌入度也可以作为关系识别的一个重要的结构性因素，因为标明了两个人的社交圈有多少是重合（overlap）的。这就给出了一种最简单的识别方法，即找出一个用户u，与之相连的节点中嵌入度最大的用户v即可认为是u的伴侣。

我们将评估基于嵌入度以及其他度量的预测方法，根据他们的表现：它们能够正确识别的案例数量。根据这一标准，嵌入度的度量达到24.7%的性能，这一方面证明了利用结构性特征进行预测的可能性，同时也提供了一个基准，用来和其他的预测方法进行比较。

接下来，我们表明，通过使用新的网络度量——离散度，可以实现两倍多的性能提升。除了想对性能较高外，其绝对精度也，举例来说，在我们的样本中，通过计算用户的离散度，其得分最高的人有60%的概率满足情侣关系。考虑到我们样本中每个用户至少有50个朋友，则他们的随机概率为2%，与此相比，我们的方法在性能上提升了30倍。

离散度的理论基础（Theoretical Basis for Dispersion）

我们测量离散度的想法首先来自于对嵌入度的作为预测依据的一点缺陷，基于社会焦点理论。许多人周围的朋友群体事实上对应于他们在生活中的焦点，例如这种群体可能是同事关系，或是大学同学。因为在这样一个群体中的许多人之间相互认识，这种小集群包含链接的嵌入度很高，即使他们之间不一定是强关系。相比之下，与一个人的伴侣或是亲密朋友相连的那条链接可能有较低的嵌入性，但他们往往会涉及共同邻居从几个不同的焦点，反映的社会轨道上的这些亲密的朋友本身之间并没有什么交集。举例来说，一个丈夫很可能认识他妻子的同事、家庭成员以及以前的同学等等，虽然这些人可能和他完全来自不同个圈子和社群。

因此,而不是嵌入性,我们建议一个个体之间的联系u和他或她的伙伴应该显示一个“分散”结构:u和v的共同邻居不是人脉广泛的,因此u和v法共同作为唯一的这些网络的不同部分之间的中介。一个例子(参见图2)。

我们现在制定一个序列捕获离散度的概念的定义。开始,我们的子图顾诱导u和所有的邻居,和一个节点在顾我们定义Cuv v的集合u和v的共同邻居中表达的观点对节点Cuv应该远顾当我们不考虑通过u和v两步路径本身,我们定义的绝对分散uv链接，disp(u，v)，是两两节点之间距离的总和在Cuv，顾以−{ u，v };这是…的dv是一个距离函数在节点Cuv。dv的函数不需要用标准距离，选择不同的dv绝对离差会产生不同的措施。下面我们详细讨论，一个大阶级中可能的距离函数，我们最终找到最好的性能定义dv(s，t)是函数s和t等于1时并没有直接联系，也没有共同邻居顾除了u和v，否则，等于0。目前的讨论中，我们将使用这个距离函数作为分散的措施的基础，下面我们考虑选择距离的影响函数。例如，在图2中，disp(u，h)= 4下这个定义和距离函数，因为有四双节点在Cuh也并不直接相关

没有共同的邻居，在顾−{ u，h }。相比之下，disp(u，b)= 1在图2中，因为一双和e形式唯一实证在幼崽没有邻居节点在顾−{ u，b }。