基于社交网络的好友推荐系统

一、问题描述

随着移动互联网浪潮的兴起，越来越多的社交网站或应用程序在网民中得到迅速普及。大多数社交应用程序都可以向用户推荐可能认识的人，例如Facebook、腾讯QQ、新浪微博、领英等。这就是诸多社交平台的重要组成部分之一——好友推荐系统。

二、问题建模

图论是适用于社交网络领域的有效建模工具之一。在社交网络的领域中，我们可以采用简单无向图或有向图进行建模。下面给出无向图的介绍。

给定一个顶点集合 $V:\{v|v\in V\}$，边集合 $E:\{\{x,y\}|(x,y)\in V^2\Lambda x\neq y\}$，二元组 $G=(V,E)$ 是一个无向图。在无向图中，节点之间存在对称的“邻接”关系。“邻接”的数学定义是：对于两个节点 x 和 y，如果无序对 $\{x,y\}$ 是一条边，则称它们是邻接的。因此，我们可以用邻接矩阵 A 来表示图 G，A 是一个 $n\times n$的矩阵，其中 n 是 G 的顶点个数，邻接矩阵 A 中的每个元素 A_{ij} 表示节点 i 和 j 之间的边的数量。显而易见，对于一个简单图，总有 $A_{ij}\in\{0,1\}$，分别表示“相连”和“不相连”，而$A_{ii}=0$（即相同的顶点之间没有边）。

对于社交网络，我们可以把每一个用户看作是一个顶点。如果两个用户之间存在好友关系，那么就在两者之间建立一条边。因此，所有用户的好友关系就可以用一张图来表示。其中，每个用户的好友数即对应图中每个顶点相连的边的条数，称作顶点的度。

与微信、领英这种双向确认的社交网络（用户 A 和 B 之间形成好友关系需要通过双方的确认）不同，对于微博、推特这类单向关注的社交网络（用户 A 可以关注用户 B 而不需要得到用户 B的允许）就需要建模成简单有向图，即边集合变成了 $E:\{(x,y)|(x,y)\in V^2\Lambda x\neq y\}$，如果有序对$(x,y)$ 是一条边，则表示用户 x 关注了用户 y。节点的出度（即以该节点为起点的边的数量）表示该用户关注的用户数，节点的入度（即以该节点为终点的边的数量）表示关注该用户的用户数。

在社交网络中，有一个最基本的原则：“在一个社交圈内，若两个人有一个共同的朋友，则这两个人在未来成为朋友的可能性就会提高。这就是著名的三元闭包原则。

这比较符合我们直觉上的想法，如果两个用户的共同好友很多，那么这两个用户互相认识并成为好友的概率就自然很大。

根据三元闭包原则，两个用户之间共同好友的数量决定了这两个人关系的强弱，所以我们可以基于两个用户之间共同好友的数量进行好友推荐。

2.1 基于共同好友数的好友推荐方法

$$Score(i,j)=|Neighbor(i)\cap Neighbor(j)|$$

上式中，$Neighbor(i)$ 表示用户 i 的好友，也就是社交网络上顶点 i 的所有邻接节点的集合。

对于有向社交网络，可以通过出度计算共同关注数，来作为好友推荐的指标：

$$Score(i,j)=|Out(i)\cap Out(j)|$$

其中，$Out(i)$ 表示用户 i 关注的用户列表，也就是社交网络上从顶点 i 出发的所有邻接节点的集合。

对于每一个用户，通过对他与其他所有用户的共同好友数进行排序，即可产生一个好友推荐列表。
由于社交网络对应的图的邻接矩阵十分稀疏，想要从图的全局信息中捕获对好友推荐有用的信息比较困难。而这种基于共同好友数量的好友推荐方法，只需要去考虑二阶连通路径的数目，其优点是计算简便，复杂度低，但是预测精度不够理想。

2.2 基于双方好友数加权的好友推荐方法

导致基于共同好友数的好友推荐方法效果较差的因素在于，如果某个用户好友人数比较少，那么无论是向该用户推荐其他好友还是将该用户推荐给其他用户都会变得很难。所以我们可以除以杰卡德系数进行加权，也就是除以双方好友总数，以减轻用户好友数差距较大的影响。

$$Score(i,j)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

2.3 基于共同好友数加权的好友推荐方法

上述的基于共同好友数的和基于加权双方好友数的两种好友推荐方法对每个共同好友的权重视为相同。但实际上在共同好友中，有些人好友多，有些好友少，当某个共同好友的好友数较少时，这个共同好友应该更加重要，所以将每个共同好友的好友数量的倒数作为系数进行加权：

$$Score(i,j)=\sum_{k\in Neighbor(i)\cap Neighbor(j)}\frac{1}{|Neighbor(k)|}$$

当然，如果不同共同好友的好友数量相差过大，可以考虑采用每个共同好友的好友数量的倒数的平方根、对数等作为系数进行加权：

$$Score(i,j)=\sum_{k\in Neighbor(i)\cap Neighbor(j)}\frac{1}{\sqrt{|Neighbor(k)|}}$$

$$Score(i,j)=\sum_{k\in Neighbor(i)\cap Neighbor(j)}\frac{1}{log_2|Neighbor(k)|}$$

2.4 基于三度影响力理论的好友推荐算法

由哈佛大学教授 Stanley Milgram 提出的六度分隔理论（英语：Six Degrees of Separation）认为世界上任何互不相识的两人，只需要很少的中间人（不超过六个人）就能够建立起联系。即只需要 6 步就可以联系任何两个互不相识的人。
由哈佛大学教授 Cristakis 提出的三度影响理论，认为我们的行为会在网络上产生波动，影响我们的朋友（一度），我们朋友的朋友（二度），以及我们朋友的朋友的朋友（三度）。如果超出三度分隔，我们的影响就会消失。

因此，在六度分隔的社会网络中，人们的行为、态度、情绪都会在三度连接以内互相影响，而在超出三度连接以外的连接则只传递信息，不再影响行为。我们称这种能对用户行为产生影响的三度分隔内的连接为强连接，三度以外的连接成为弱连接，只能传递信息，不能影响好友行为。

所以为了达到比仅考虑二阶连通节点的基于共同好友的好友推荐算法更好的预测性能，出现了基于三度影响力理论的好友推荐算法。

定义用户 x 和用户 y 之间的好友关系强度计算公式如下：

$$R(x,y)=\sum_{i=2}^{3}\frac{1}{i-1}\cdot\frac{N_{x,y}^{i}}{\prod_{j=2}^{i}(n-j)}$$

其中，$1/(i-1)$ 是随路径长度增加而减少的权重，即路径长度越长，好友关系强度越小。n 表示局部社交网络图 G 中的节点总数量，$N_{x,y}^{i}$ 是图 G 中实际存在的连接节点 x 和 y 长度为 i 的连通路径的数量，$\prod_{j=2}^{i}(n-j)=(n-2)*(n-3)*\ldots*(n-i)=A_{n-2}^{i-1}$ 是假设图 G 中任意两节点都有边相连时，可能存在的连接节点 x 和 y 且长度等于 i 的所有路径的数量，而长度为 i 的所有可能路径的数量是从除 x 和 y 外的其他 $n-2$ 个节点中选取 $i-1$ 个节点连接所构成的全排列。由于算法只考虑距离 3 以内的强连接关系，所以 i 的取值为 2 或者 3（距离为 1 的用户已经是好友了，无需再进行好友推荐）。

2.5 融合时间特征的基于共同好友数加权的好友推荐算法

在上述的所有方法中，都简单地将社交网络抽象为了静态图，即没有考虑不同时刻下的社交网络以及网络中节点间的联系是具有时序性的。
而 Facebook 则在其好友推荐系统中考虑了时序信息，这里简单介绍该系统的思路。

该系统主要基于以下这个假设：用户对新添加的好友更感兴趣。例如，f_1, f_2 都是用户 u 的好友，f_1 是近期添加的好友，f_2 则是很久之前添加的好友，则用户 u 对今天添加的好友 f_1 更感兴趣。
基于该假设，可以得出如下的经验公式：

$$v(target)=\sum_{k\in Neighbor(u)}\frac{(\delta_{u,k}\cdot\delta_{k,target})^{-0.3}}{\sqrt{|Neighbor_k|}}$$

可以看出，该公式只是在共同好友数加权的公式上增加了时间特征 $(\delta_{u,k}\cdot\delta_{k,target})^{-0.3}$。
其中，$\delta_{u,k}$ 为 u 和 k 建立好友关系至今经过的时间，$\delta_{k,target}$ 为 k 和 $target$ 建立好友关系至今经过的时间，-0.3 为惩罚因子，为 Facebook 根据实际业务情况设置的经验参数。从该公式不难看出，时间相差越大，权重越小。

根据以上经验公式，我们可以直接计算得出所有潜在的推荐候选人的好友推荐得分，作为系统进行好友推荐的重要依据或输入特征之一。

三、总结

本文将社交网络建模为有向图或无向图，详细介绍了基于共同好友数的好友推荐算法、基于三度影响力理论的好友推荐算法和融合时间特征的基于共同好友数加权的好友推荐算法，分析了这些算法所以依赖的假设及社会学依据，指出了这些算法为好友推荐系统决策所提供的指标计算公式，最后讨论了每个算法存在的一些缺点和不足。