毕设进度简述 20110120

3071102314 周晓龙
frogcherry@gmail.com
13675870206

课题：动态社会网络挖掘

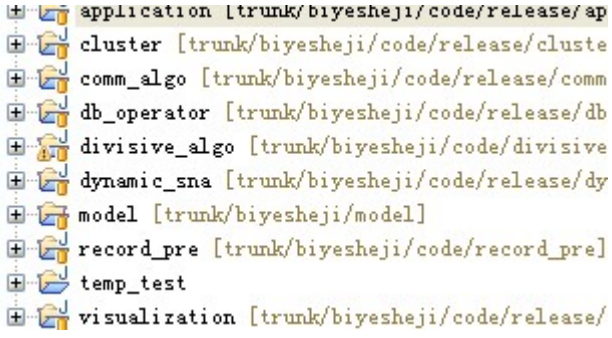
基本研究思路概要：

1) 动态网络挖掘，我们的定性为分析在以时间维度为动态线，网络中团体的演变过程和个人的活动过程。
2) 以时间维度为基线，一个最行之有效的方法就是对动态网络的在各个时间点上的状态做切片，然后分析各切片之间的联系和演变
3) 从研究目标可知，研究的两个基本点即重点为：其一，单时间切片上的网络团体发现；其二，时间片之间的团体演化分析和个人活动分析
4) 对于第一点，我们采用一个聚类分析的过程进行团体发现。准备采用和现阶段已经实现的算法是，Michelle Girvan 和 Mark Newman 提出的 betweenness 切边算法。算法基本思路比较简单，网络中的关键路径是团伙之间的联系，通过去掉网络中的关键路径，即可以发现团伙。
5) Betweenness 切边算法的关键就落在关键路径的查询上了。经典算法中采用各个点两两之间的最短路径的叠加来寻找关键路径。我采用 Ulrik Brandes 在 A Faster Algorithm for Betweenness Centrality 一文中提出的一个优化的 betweenness 计算方式，细节比较复杂这里就不具体展开了。
6) 有了基本切片算法，还不足够。切片算法有个与生俱来的缺陷，就是需要事先指定一个切去的边的条数。另外，对于聚类分析的团伙结果，我们也需要做一个质量检测。因此，我们还需要一个检测团伙聚类质量的度量。M. E. J. Newman 和 M. Girvan 在其经典论文 Finding and evaluating community structure in networks 中就提出一个聚类效果的度量：Modularity Qualify。我现阶段已经实现了这个算法。
7) 有了度量，我们可以对切边算法做进一步优化，使其自动化起来。通常来说，逐一切片，网络的 MQ 度量会呈现一个或多个峰值，最高峰值是理论上的最优聚类结果。我们对网络逐一切片，记录过程中的切边过程和 MQ 度量。以实现两个方式发现最优结果：a.贪心法找到第一个峰值结束；b.逐一切去所有边，找到最高峰。另外切边过程中记录过程值，可供使用者手工调整达到最优。
8) 至此，第一个基础点，单切片上的团体发现就有一个比较完善的方案。下面是动态的分析。
9) 时间维度上的团体演化，可以分为以下几个基本动作:1.团伙出现；2.团伙瓦解；3.团伙分裂；4.团伙合并；5.团伙保持。对于个人来说，两个基本动作：1..进入团伙；2.离开团体
10) 个人层面的演化比较容易发现，进出团体的过程

11) 团伙层面的演化的研究，难点之一在于切片之间团体的对应关系的发现和保持。简单解释就是在切片 t 上的团体 A0，到了切片 t+1 上，我们要在分出的团体中找到和 A0 对应的团体 A1。如果发生的是分裂和合并还需要追踪到最为相近的一列表。
12) 追踪的过程比较复杂也是动态挖掘的难点所在。研究的思路还是要化纷繁为简单，抓住核心。首先最基础的一个基线是找到最相近的团体。
13) 首先是一个相近度的度量，我们使用 Jaccard 系数进行度量即 C1，C2 的相似度 = |C1 交 C@| / |C1 并 C2|
14) 切片 t 中团体集合 St0={P0,P1,P2,…,Pn}，切片 t+1 中团体集合 St1={Q0,Q1,Q2,…,Qm}，两两之间(Pi 与 Qj 间)都会有一个相似度 Jaccard 系数。对于单团，取出最大的相似度对应团是非常自然的想法。但是不能这样简单地处理，必须全局地看待这个问题。即取出全局上最优的匹配组合。这样，这个问题可以转化为二部图的最佳匹配问题。
15) 对于二部图的最佳匹配问题，我采用并已经实现了 Kuhn－Munkras 算法来解决二部图匹配问题。并查找了很多资料进行了一些优化
16) 至此，侦测团体的保持和研究保持团体的演化以及团体的诞生和瓦解是比较有依据了，但是对于侦测和研究团体的分裂和合并还是有难度
17) 如果是完全的分裂和合并侦测还是比较简单的。例如分裂，P1 团体的下一时间片完全变为 Q1,Q2，没有新成员进入，也没有成员离开。但是对于比较复杂的变化，就需要定义一个值去界定分裂和合并。我们已经阅读了相关的一些论文。
18) 下一步的工作首先是把已经基本实现的算法框架可视化，再就是寻找数据集和使用数据集进行实验比对，再就是对比较复杂的团体行为进行进一步的研究和分析。

已完成工作

在上文中已经介绍过。基本算法已经实现，可视化已经完成部分。



Comm_algo 包是公共算法，包含基础的矩阵表示及计算，图的表现，一些基础数据及算法如堆等等

Db_operator 是数据库接口，包含数据表脚本，数据实体映射，数据操作类等

Model 包里面是一些软件的架构和设计，以及数据建模

Record_pre 包含一些数据的基础准备过程

Cluster 包包含一些聚类算法，主要是 betweenness 切边算法以及度量等

Dynamic_sna 包含动态网络分析的相关算法

Visualization 是可视化的实现

Application 是整合的应用 Demo

统计结果

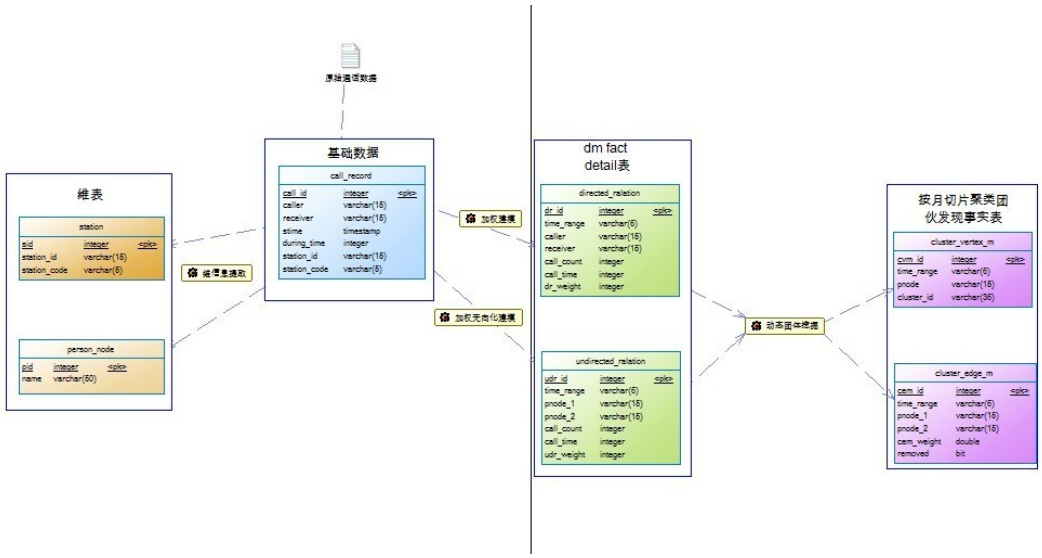
File	Type	Folder	Lines	Code lines	Code//Co...	Comment...	Blank lines
DirectedRa...	.xml	D:\study\...	30	29	0	1	0
PersonNod...	.xml	D:\study\...	15	14	0	1	0
RelationW...	.xml	D:\study\...	20	20	0	0	0
Station.hb...	.xml	D:\study\...	18	17	0	1	0
Undirected...	.xml	D:\study\...	30	29	0	1	0
DynamicCl...	.java	D:\study\...	58	25	5	18	10
ClusterSlic...	.java	D:\study\...	87	42	4	22	19
ClusterSimil...	.java	D:\study\...	43	20	2	13	8
KM.java	.java	D:\study\...	87	43	1	0	43

代码行数:	9964	总单元测试用例数:	502.25	总文件数:	221
注释行数:	3941	总单元测试缺陷数:	100.45	总文件大小(KB):	406
空行数:	1908	总结合测试用例数:	200.90	总人月数:	2.51
代码//注释:	81	总结合测试缺陷数:	20.09	总成本($):	25112.50
总行数:	15894				

保存...(V)

如上，已实现软件规模约 15894 行。主要平台采用 Java

该数据集市的基本建模如下所示，已准备数据是通话记录，以后的研究可能会使用到其他的数据集。



可视化的实现技术我并不熟悉，正在边学边做，已经做的是一个简陋的画出相邻两个切片的界面：
相同颜色是找出的最佳匹配团。