数据挖掘第一次作业

杨一凡　1300010703　　郭宇航　1300010726

1 问题的分析

这次的作业使用的数据集是DBLP中关于论文合作关系的数据，主要有四个挖掘任务。

任务一：给挖掘出的合作关系赋予有意义的权值，来表示合作的紧密性。
看到问题的第一想法就是挖掘频繁二项集，运用两个作者的合作文章的数量来表示两个作者合作紧密程度的度量。当然，可能存在一些改进，我们可以假设如果一篇文章的合作人数较多，那么这篇文章对合作作者的联系贡献就会比较少。基于这一假设，我们可以将每一篇文章除以作者数n或者C_n^2加总到作者的联系矩阵中。

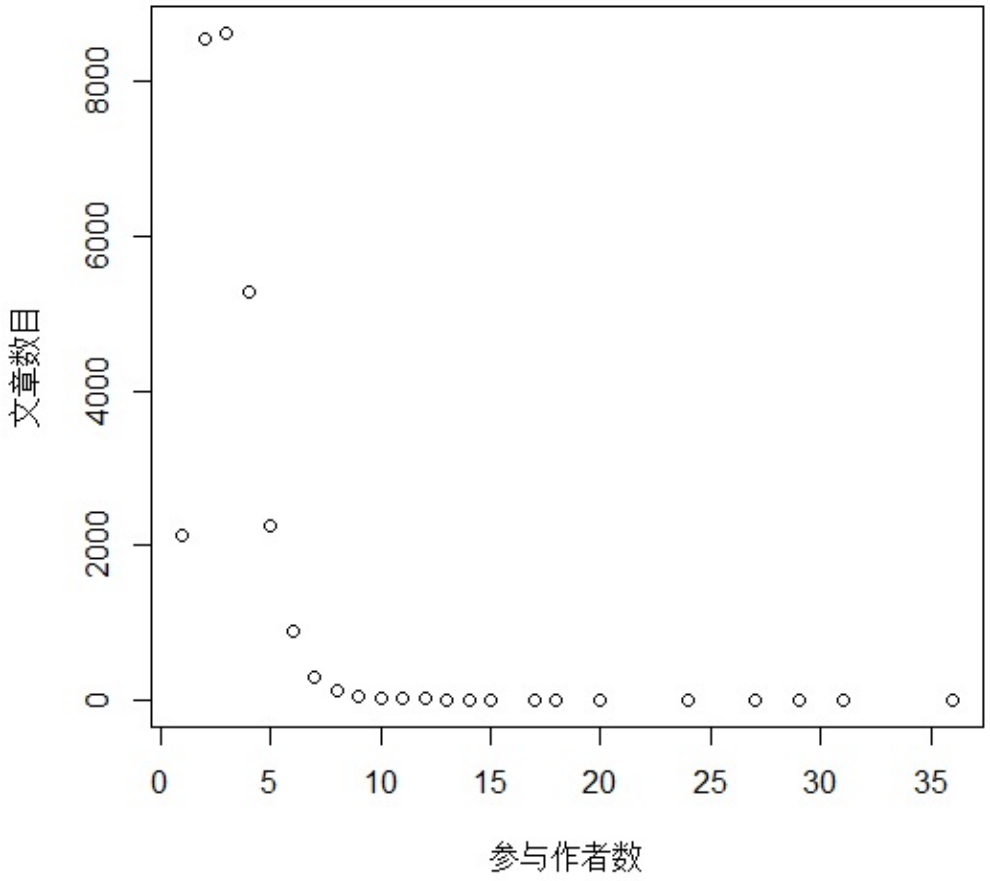
任务二：挖掘导师-学生指导关系和指导时间，并进行一定正确性的分析和验证。
一个很自然的假设是导师-学生关系应该是作者紧密合作关系的子集，因此我们期望从任务一的联系矩阵中取得那些具有紧密合作关系的作者，再从中获取师生关系。一般来看，导师的年龄一般会比学生的年龄大，因此我们将具有紧密合作关系而活跃时间具有一定差异的合作关系判定为师生关系。

任务三：挖掘频繁合作关系（如果有多人经常在一起合作，请挖掘出合作团队）。
我们理解这一个任务的目的是挖掘闭的频繁模式，并且那些长度大于一的闭频繁模式会比较有意义。我们打算用FP树来挖掘闭的频繁模式。

任务四：挖掘各个会议的“核心”研究者。
很自然的想法是运用作者在各个会议发表文章的数量来衡量作者的活跃程度。但我们觉得仅仅运用会议论文发表数量来衡量作者的活跃程度太过于单一了，因此，我们打算在结合作者的PR分值(根据PageRank算法获得的作者打分)来评估作者的活跃程度。

2 数据的预处理

因为这次作业针对的是会议相关的文献，并且只需要作者、年份和会议的信息，我们最先提取了dplp.xml中标注为proceedings和inproceedings的文章中的keys、author和year的信息(保存在文件data1.txt)。由于题目中要求获取2000 年以后的数据，我们进一步筛选data1.txt，筛选出发表时间在2000 年以后的数据(保存在new_data 中)。从拥有不同数量的文章数量来看，大多数的文章是由两到三个作者合著完成。针对不同的任务，我们对数据进行了不同处理，为了挖掘各个会议相关的信息，我们将new_data 进行了拆分，分别保存在“会议名.txt”中；为了简化在FP 树算法中的处理，我们仅提取作者的信息，将其保存在final_data.txt中；为了获取作者的活跃时间，我们将每个作者发文献的年份的平均值作为每个作者活跃时间。


文章数目
0
2000
4000
6000
8000
0
5
10
15
20
25
30
35
参与作者数

3 PageRank算法

由于这次作业的其他处理或者是比较初等的数据处理，或者是上课已经讲过的FP树算法，就不进行分析了。这一小节主要对PageRank 算法进行讨论。

PageRank算法是谷歌创始人Larry Page提出的对网页进行打分的算法。基本思路是高评分的网页其链出的网页会受到这个高评分网站的影响，因而具有高评分。一个网页对其链出网页的贡献会因为链出的网页数量增多而被稀释。事实上，PageRank算法不只是可以用于网页的打分，基于一个带权有向图，PageRank算法都可以获得有向图中点的得分。因为我们的数据是关于合作关系的(如果是引用关系则可以建立带权有向图)，我们可以建立一个带权无向图，运用PageRank算法对图中的点进行打分，下面给出PageRank算法在这次作业中的算法框架。

第一步：从new_data中获取作者节点的连接关系与度数，用n_{ij}来表示点i和j之间连接的边的数量(也就是两个作者合作文章的数量)，用d_i表示点i的度数(即是某一作者与其他作者合作的次数)。
第二步：初始化作者得分PR^0，在page_rank函数中，我们将所有作者的初始得分记为1。
第三步：迭代。迭代规则如下

$$PR^{k+1}(i) = \sum_{j=1}^{N} \frac{PR^k(j) \cdot n_{ij}}{d_j}$$

如果PR^{k+1}与PR^k的差距足够小就退出迭代；否则，重复第三步。

经过以上迭代，我们便会获得相关作者的PageRank得分。因为通过迭代获得的PageRank得分，我们不禁疑惑通过如是迭代获得的结果是否存在，是否唯一。下面，我们将分析迭代结果的存在唯一性。

其实PageRank算法的模型是假设作者在网页链接形成的有向图上进行随机游动，PageRank得分也就是这个随机游动的一个不变测度(不变分布)。因为随机游动是一个马氏过程，我们可以写出马氏过程的转移矩阵

$$T = \begin{bmatrix} 0 & n_{12}/d_1 & n_{13}/d_1 & \cdots & n_{1N}/d_1 \\ n_{21}/d_2 & 0 & n_{23}/d_2 & \cdots & n_{2N}/d_2 \\ n_{31}/d_3 & n_{32}/d_3 & 0 & \cdots & n_{3N}/d_3 \\ \vdots & \vdots & \vdots & & \vdots \\ n_{N1}/d_N & n_{N2}/d_N & n_{N3}/d_N & \cdots & 0 \end{bmatrix}$$

事实上，这一个转移矩阵即是我们进行迭代的矩阵。换句话说，PageRank的迭代形式即是

$$PR^{k+1} = PR^k \times T$$

我们有以下定理保证迭代的收敛性

Perron-Frobenius 定理 设$A = (a_{ij})$为$n \times n$不可约实矩阵，所有元素均非负，即$a_{ij} \geq 0$，那么下列结论成立：
1.存在一个实的特征值r，其他特征值λ的模均不超过r，即$|\lambda| \leq r$;
2.与r对应的特征向量的所有元素均非负;
3.$min_i \sum_j a_{ij} \leq r \leq max_i \sum_j a_{ij}$;

转移矩阵T的每一行元素的和为1，即$\sum_{j=1}^{N} n_{ij}/d_i = 1$，由以上定理的第一条和第三条可知转移矩阵的最大特征值为1，因此我们通过迭代可以让$PageRank$得分收敛到转移矩阵特征值为1的特征子空间，并且收敛速度可以用一个几何级数控制。

在随机过程中，我们有定理：
不可约、常返的转移矩阵的不变测度在忽略常数倍的意义下唯一。
而有限图上随机游动的常返性等价于图的连通性。所以，只要合作关系对应的带权无向图是连通图，我们就能保证PageRank分值在相差一个常数倍的意义下唯一。

4 数据挖掘结果

4.1 任务一

我们运用字典来保存合作关系，以一个python的二元元组作为键，以二元组的总分键值，元组按字典序排序解决顺序问题。在第一小节提出了每个二元组出现的分值为1，或者与参与论文的作者数量有关，这里我们取系数为$1/C_N^2$，在接下来的表中，将列出在两种计量方式下合作关系密切的top5，其余信息保存在文件relation.txt和relation_0.txt中。

合作关系	合作关系得分
Craig Macdonald,Iadh Ounis	32
Jun Yan,Zheng Chen	30
Irwin King,Michael R. Lyu	25
Dinh Q. Phung,Svetha Venkatesh	25
Qiang Yang,Zheng Chen	25

Table 1: 第一种计分方式合作关系top5

合作关系	合作关系得分
Craig Macdonald,Iadh Ounis	7.67
Pascal Fua,Vincent Lepetit	5,63
Irwin King,Michael R. Lyu	5.33
Corinna Cortes,Mehryar Mohri	5.20
Claudio Gentile,Nicol ò Cesa-Bianchi	5.17

Table 2: 第二种计分方式合作关系top5

4.2 任务二

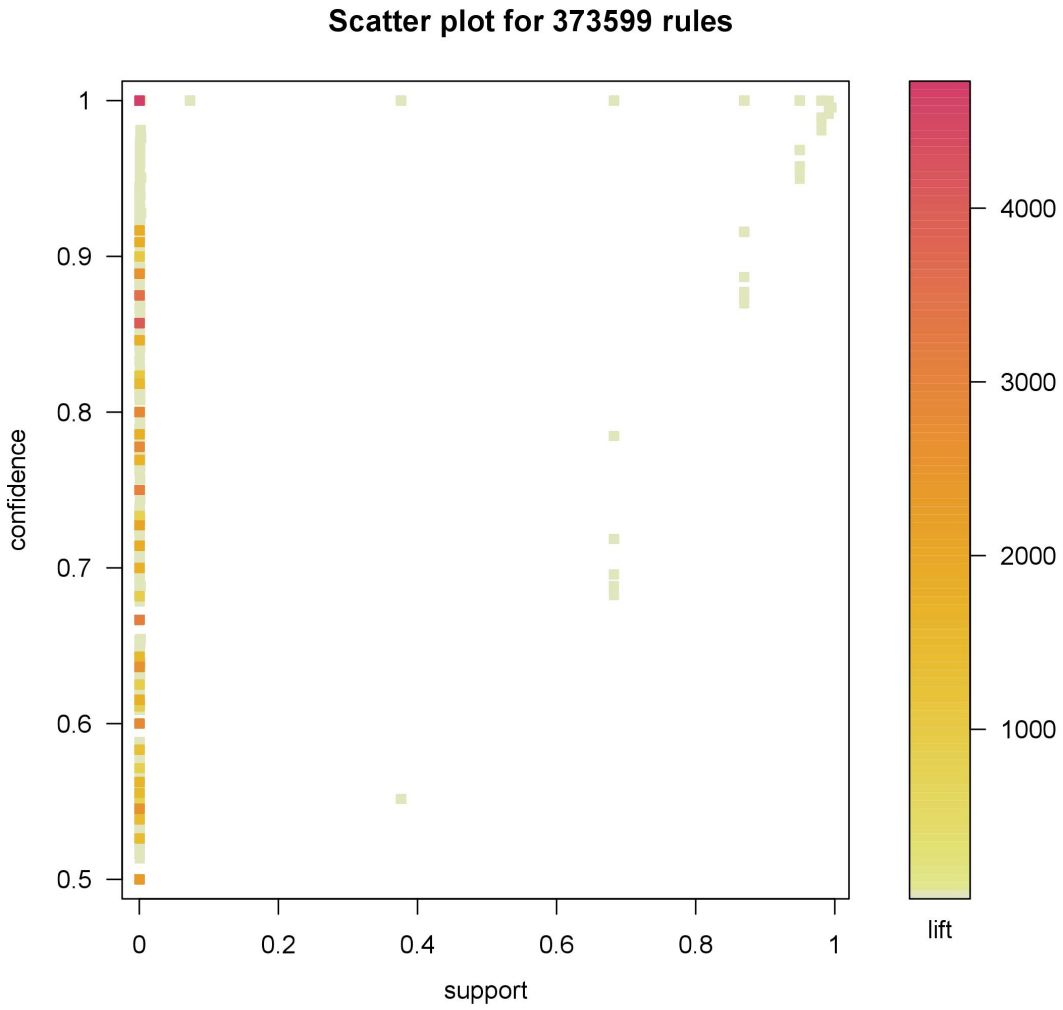
根据第一小节的叙述，我们选出两人平局活跃年份差距在10年以上，根据学生导师的关系得分，取出最高的top10。

4.3 任务三

我们先调用了R语言的arules包，获得了数据集频繁项支持度和置信度的关系，后面自己用python语言编写了FPtree算法。我们挖掘所有的支持度超过20 频繁闭模式，从挖掘的结果来看大多数的频繁项集是单项集见图，因此我们在表中列出非单项集的频繁项集所有的频繁闭模式。

老师	学生
Xiaoyang Sean Wang	Zhen He
Xiaolin Feng	Yang Song
Thomas M. Mann	Siegfried Handschuh
Timothy Hancock	Peter Christen
Ole Moller Nielsen	Peter Christen
Mikhail J. Atallah	Yasin N. Silva
Mikhail J. Atallah	Ruby Y. Tahboub
Mikhail J. Atallah	Qutaibah M. Malluhi
Mikhail J. Atallah	MingJie Tang
Mijail Serruya	Wei Wu

Table 3: 老师学生关系



Shusaku Tsumoto,Shoji Hirano
Xueqi Cheng,Jiafeng Guo
Philip S. Yu,Xiangnan Kong
Chun Chen,Jiajun Bu
Shiguang Shan,Xilin Chen
Svetha Venkatesh,Dinh Q. Phung
Mehryar Mohri,Corinna Cortes
Pascal Fua,Vincent Lepetit
Michael R. Lyu,Irwin King
Xiaofei He,Deng Cai
Nicol ò Cesa-Bianchi,Claudio Gentile
Zheng Chen,Jun Yan
Iadh Ounis,Craig Macdonald
Philip S. Yu,Wei Fan
Qiang Yang 0001,Zheng Chen
Philip S. Yu,Charu C. Aggarwal

Table 4: 支持度超过20的频繁闭模式

4.4 任务四

我们打算用作者在每个会议的PageRank分值作为主要评价指标，挖掘结果列在“result 会议名.txt”的文件中。下表列出各个会议的影响力前三的文章。

5 尚存在的问题

对于两人合作关系的评价指标，过于单一，可能的改进是用作者本人的一些特征来改进单一的两人合作次数；师生关系的推测，太直接了，不过没有想出太好的办法；在FPtree的实现中，是使用递归调用来实现的，并不能保证达到最高的效率；至于PageRank算法，其分值的可解释性太弱，虽然好用，到难以有令人信服的解释。

会议	top 1	top 2	top 3
colt	Manfred K. Warmuth	Peter L. Bartlett	Maria-Florina Balcan
dmkd	Zheng Chen	Maarten de Rijke	Marcos Andr é Gongalves
cvpr	Shuicheng Yan	Thomas S. Huang	Xiaoou Tang
icml	Lawrence Carin	Bernhard Scholkopf	Michael I. Jordan
icdm	Jiawei Han	Philip S. Yu	Hui Xiong
pakdd	Christos Faloutsos	Joshua Zhexue Huang	Longbing Cao
nips	Bernhard Scholkopf	Andrew Y. Ng	Yoshua Bengio
wsdm	Yi Chang	Vanja Josifovski	Xueqi Cheng
tkde	Jiawei Han	Philip S. Yu	Ming-Syan Chen
sigir	Jaap Kamps	Nicola Ferro	Mark Sanderson
sdm	Philip S. Yu	Jiawei Han	Vipin Kumar

Table 5: 会议核心作者top3

参考文献：

[1]钱敏平. (2011). 应用随机过程. 北京: 高等教育出版社.

[2]Langville, A. N., Meyer, C. D., & Books24x7, I. (2006;2011;2009;2012;). Google's pagerank and beyond: The science of search engine rankings. Princeton, NJ: Princeton University Press.