# 机器学习第三次作业

目录

[机器学习第三次作业 1](#_Toc502785908)

[数据预处理 1](#_Toc502785909)

[第一题第一问 2](#_Toc502785910)

[第一题第二问 4](#_Toc502785911)

[第二题第一问 5](#_Toc502785912)

[第二题第二问 7](#_Toc502785913)

[总结 9](#_Toc502785914)

## 数据预处理

这次作业的要求是针对DBLP数据库中IJCAI, AAAI, COLT, CVPR, NIPS, KR, SIGIR, SIGKDD八个会议中从2007 年至今的所有数据进行处理，因此在进行实验之前，首先要对数据进行预处理，避免一些不正确的数据影响结果。例如：

1. 去掉没有作者的条目，例如《IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013》这样属于会议主题的条目；
2. 去掉不在题设中八个会议的论文；
3. 对于类似MPREF@AAAI这样的子会议，全部归为其父会议，比如，MPREF@AAAI替换成AAAI；值得注意的是，由于数据集中没有SIGKDD，因此用KDD代替；
4. 作者名称中可能出现数字标号，例如“Jiawei Han 0001”，将数字标号去除；
5. 去掉会议前言这样没有主题的论文。

然后将数据按照如代码1所示的数据结构进行存储，方便之后用于实验。

代码1 数据集数据结构

{  
 'conference': xxx,  
 'title': xxx,  
 'year': 20xx,  
 'author': [  
 xxx,  
 xxx,  
 xxx  
 ]  
}

## 第一题第一问

第一问要求找到各会议的研究者。基于已经进行预处理过的数据集，我们将数据集中各论文的作者以及作者最新发表的论文提取出来，并按照会议和时间分类，得到各会议根据时间分类的研究者列表。然后根据时间就可以分析哪些研究者依旧活跃，哪些研究者已经不再活跃。

在这里我们定义，作者最新发表的论文的时间在2013年以前（包含2013年）的认为是不活跃的，也就是说作者自2013之后就再也没发过论文。那么对应地，作者最新发表论文的时间在2013年之后的就认为是活跃的了。

以CVPR会议为例，首先我们可以看到，对于CVPR会议来说，其研究者有：

{2007: ['Mayumi Yuasa',

'Tomokazu Wakasugi',

'Lech Szumilas',

'Rene Donner',

'Helen M. Meng',

'Vincent Cheung',

'Suneel Suresh',

'Andy Tsai',

'Alfred O. Hero III',

'Niranjan Damera-Venkata',

'Zhen Wen',

'Mika Fischer',

'Qin Jin',

'Wai Han Ho',

'Paul Andrew Watters',

……

}

（详见homework-3.ipynb，以下同）

然后我们找出CVPR的活跃研究者，也就是发表时间在2014年以及之后的作者：

{2014: ['Armand Joulin',

'Enrique G. Ortiz',

'Sanath Narayan',

'Mohan S. Kankanhalli',

'Kalpathi Ramakrishnan',

'Wei Li ',

'Richard V. Stebbing',

'Olof Enqvist',

……

}

同样我们也可以找到所有CVPR不再活跃的研究者：

{2007: ['Mayumi Yuasa',

'Tomokazu Wakasugi',

'Lech Szumilas',

'Rene Donner',

'Helen M. Meng',

'Vincent Cheung',

'Suneel Suresh',

'Andy Tsai',

'Alfred O. Hero III',

……

}

对于CVPR来说，其研究者最后发表时间的分布如下图所示：

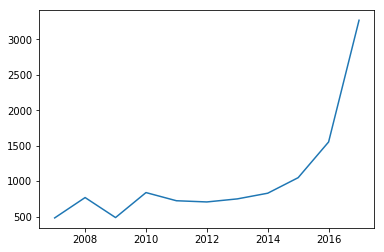


图2-1 CVPR的各研究者最后发表的论文其时间分布

可以看到，大部分作者还是比较活跃的，我们再统计一下活跃/不活跃的作者数量，发现活跃的作者有6698人，不活跃的作者有4758人，目前仍比较活跃的作者占了一半以上。

图2-2 CVPR活跃/不活跃的研究者占比

## 第一题第二问

第二问要求我们找到经常在一起合作的学者，首先我们通过频繁模式挖掘可以找到经常一起发表论文的作者组。在课上所讲授的频繁模式挖掘的实现有两种，一种是Apriori，另一种是FPGrowth。Apriori基于先验性质：频繁项集的非空子集也一定是频繁的，对候选项集逐层搜索，直到找不到频繁项集为止。Apriori的缺点在于会产生大量的候选项集，且需要对所有候选项集进行计算，因此算法的时间复杂度和空间复杂度较高。FPGrowth则是不产生候选项集，直接找到频繁项集。首先由长度为1的频繁模式开始，构造条件模式基，然后基于条件模式基构造条件FP树，并递归地在该树上进行挖掘，最终将后缀模式和条件FP树产生的频繁模式连接得到频繁项集。可以看到FPGrowth算法的复杂度要低于Apriori，在实践中也是如此。数据集经过预处理后共有26694条数据，处理这些数据FPGrowth只需要3-4分钟，而Apriori则需要半个小时以上。

经过FPGrowth处理后我们可以得到所有频繁项集。在这里我们根据题目要求，将“经常在一起合作”定义为最小支持度为3，项集的规模至少为3人。经过筛选我们得到所有符合标准的频繁项集：

['Lei Wang ', 'Xinwang Liu', 'En Zhu'] 3

['Lei Wang ', 'Yong Dou', 'En Zhu'] 3

['Lei Wang ', 'Xinwang Liu', 'Yong Dou', 'En Zhu'] 3

['Lei Wang ', 'Jianping Yin', 'En Zhu'] 3

……

……

['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu'] 12

['Feiping Nie', 'Heng Huang', 'Chris H. Q. Ding'] 13

['Feiping Nie', 'Heng Huang', 'Hua Wang'] 15

['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan'] 18

['Yiqun Liu', 'Min Zhang ', 'Shaoping Ma'] 18

但是分析这些频繁项集就会发现，其中会出现[A, B, C], [A, B, C, D]以及[A, B, D]这样的团队，为了解决这一问题，我们对频繁项集进行处理，将关联度较大的频繁项集合并在一起。首先我们处理[A, B, C]和[A, B, C, D]这样的情况，我们按照支持度从高到低，将频繁项集与其子集合并，即将[A, B, C, D]合并到[A, B, C]下面，[A, B, C, D, E]合并到[A, B, C, D]下面。然后我们将[A, B, C]下的所有子集里的作者提取出来，这样就能得到[A, B, C, D, E]这样的集合，在这里我们称为Author List。可知这样得到的Author List很可能出现重复，因此我们将所有相同的Author List以及具有包含关系的Author List合并在一起，将所有相关人员聚合，最终得到团队。结果如下所示：

Group Member: ['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu', 'Shengxian Wan', 'Liang Pang', 'Shuzi Niu', 'Long Xia']

Contains:

Author List:['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu', 'Shengxian Wan', 'Liang Pang', 'Shuzi Niu', 'Long Xia']

Group: ['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan']

Subgroup:[

Group: ['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu']

Subgroup:[

Group: ['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu', 'Shengxian Wan']

Subgroup:[

Group: ['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu', 'Shengxian Wan', 'Liang Pang']

Subgroup:[

]

]

……

（详见result\_1.txt）

这样我们就能得到1001组团队，其形式如代码中Group Member所示，包含了所有合作过的研究者。结果中Group Member是团队成员，Contains包含了所有组成该团队的频繁项集，Group即父频繁项集，Subgroup即其子频繁项集（父频繁项集是子频繁项集的子集）。

## 第二题第一问

根据每个团队发表论文的情况，提炼出该团队最常涉猎的主题词。我们利用LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）对标题集进行学习。LDA是一个无监督学习算法，在训练时只需要指定文档集和主题数量k即可。我们将所有的标题提取出来，然后训练LDA模型。在训练时要对训练集进行预处理，首先将标题切分成短语，即Tokenize；然后去除停用词，即高频没有什么意义的词语，类似at、or这样的介词或者连接词；接下来将词语归一成一个形态，例如将plans变成plan，将learning变成learn等等；最后将词语汇总成词典（Dictionary）并处理成词向量，交由LDA模型进行训练。在这里我们有两种方案，一是由于不同会议的主题不同，例如AAAI是人工智能相关，CVPR是计算机视觉和模式识别相关，因此对不同会议的论文进行训练，然后根据会议不同对论文标题进行主题预测；二是直接将所有标题作为数据集进行训练并预测。可以看到，前者的预测主题的契合度明显要好于后者，但是由于有8个会议，主题数量较多，不好定主题词，因此这里我们用后一种方案，取k=30，即30个主题。在训练后得到的结果如下所示：

[(0,

'0.092\*"plan" + 0.072\*"constraint" + 0.059\*"game" + 0.056\*"agent" + 0.053\*"multi" + 0.031\*"class" + 0.028\*"new" + 0.025\*"strategi" + 0.023\*"gener" + 0.020\*"complex"'),

(1,

'0.078\*"adapt" + 0.068\*"label" + 0.060\*"domain" + 0.055\*"recommend" + 0.039\*"filter" + 0.038\*"high" + 0.037\*"context" + 0.037\*"multi" + 0.031\*"awar" + 0.030\*"dimension"'),

(2,

'0.093\*"learn" + 0.065\*"supervis" + 0.056\*"scene" + 0.054\*"view" + 0.048\*"discrimin" + 0.043\*"languag" + 0.042\*"reconstruct" + 0.041\*"machin" + 0.039\*"semi" + 0.029\*"stereo"'),

……

]

（详细见homework-3.ipynb）

共30个主题。这些主题由主题词和其概率组成，概率越大的词语，越能够诠释主题的内容。当我们用LDA模型对论文标题的主题进行预测时，也会得到一个概率集合：

[(2, 0.12916666), (9, 0.12916668), (14, 0.12916666), (18, 0.12916666), (26, 0.37916666)]

这个概率集合说明了论文标题可能的主题，例如本例，标题为《Understanding Behaviors that Lead to Purchasing: A Case Study of Pinterest.》，预测得到的结果如上所示，第26个主题的概率较大，第26个主题为：

(26,

'0.107\*"graph" + 0.083\*"rank" + 0.046\*"regular" + 0.038\*"low" + 0.033\*"bound" + 0.032\*"project" + 0.030\*"tensor" + 0.025\*"studi" + 0.023\*"bandit" + 0.023\*"base"')

为图相关的论文，不过可以看到，由于论文标题长度较短，因此使用LDA模型训练效果并不是特别好，且对于主题，主题词之间关联度不是很高，因此不好总结主题词。

训练完LDA模型后，接下来要根据每个团队发表的论文，提炼出该团队最常涉猎的主题。首先我们回到上题中的团队（Group），针对每一个频繁项集，找到该频繁项集中作者共同发表的论文，然后合并起来得到该团队发表的论文（Paper List），再利用LDA模型预测论文的主题，将所有概率模型累加起来，得到最终该团队涉猎主题的概率，取最高的前三个主题作为该团队的主要主题。运算的结果存储至result\_2.txt中。我们以['Xueqi Cheng', 'Jiafeng Guo', 'Yanyan Lan', 'Jun Xu', 'Shengxian Wan', 'Liang Pang', 'Shuzi Niu', 'Long Xia']这个团队为例进行探究。可以看到该团队在2007年-2017年间共发表了18篇论文：

Reinforcement Learning to Rank with Markov Decision Process. : 2017

Adapting Markov Decision Process for Search Result Diversification. : 2017

A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. : 2016

Text Matching as Image Recognition. : 2016

SPAN: Understanding a Question with Its Support Answers. : 2016

……

（详见result\_2.txt）

该团队的主题概率为：

{

12: 2.36863

11: 1.82403

1: 1.65095

9: 1.59895

26: 1.17301

22: 0.983603

……

}

（详见result\_2.txt）

我们从主题列表中找到前三个主题：

(12,

'0.278\*"learn" + 0.073\*"deep" + 0.041\*"represent" + 0.040\*"activ" + 0.029\*"random" + 0.027\*"process" + 0.027\*"task" + 0.024\*"multi" + 0.023\*"transfer" + 0.022\*"model"')

(11,

'0.071\*"queri" + 0.061\*"distribut" + 0.046\*"map" + 0.043\*"prefer" + 0.039\*"rule" + 0.038\*"data" + 0.033\*"constrain" + 0.030\*"driven" + 0.029\*"analysi" + 0.027\*"complet"')

(1,

'0.078\*"adapt" + 0.068\*"label" + 0.060\*"domain" + 0.055\*"recommend" + 0.039\*"filter" + 0.038\*"high" + 0.037\*"context" + 0.037\*"multi" + 0.031\*"awar" + 0.030\*"dimension"')

即深度学习、搜索分析以及推荐系统这三个主题。

## 第二题第二问

接下来我们要探究，团队和团队所对应的主题随着时间的变化。考虑到数据集的规模，我们设5年为变化周期，即针对2007-2012年（包括2012年）的论文建立团队，然后针对2013年-2017年（包括2017年）的论文建立团队，针对两个年代的团队，查找相似度最高的团队，可以认为，这两个团队在这两个时间段内是一个团队，然后我们对比两个团队的成员和主题即可得到团队的构成状况及其研究主题的变化情况。所有的对比结果存在homework-3.ipynb中。我们以['Wen Gao ', 'Xilin Chen', 'Shiguang Shan', 'Yu Su']这个组为例子。

在2007年-2012年，团队成员为：

['Wen Gao ', 'Xilin Chen', 'Shiguang Shan', 'Yu Su']

论文涉及的主题为：

{

17 : 1.65842,

13 : 0.755761,

5 : 0.688225,

21 : 0.685524,

12 : 0.606029,

……

}

前三个主题为：

(17,

'0.135\*"recognit" + 0.082\*"3d" + 0.068\*"action" + 0.061\*"face" + 0.054\*"improv" + 0.040\*"use" + 0.038\*"space" + 0.025\*"train" + 0.025\*"prior" + 0.022\*"model"')

(13,

'0.089\*"logic" + 0.080\*"knowledg" + 0.063\*"tempor" + 0.056\*"reason" + 0.039\*"base" + 0.039\*"descript" + 0.034\*"order" + 0.032\*"ontolog" + 0.024\*"s" + 0.023\*"program"')

(5,

'0.228\*"network" + 0.091\*"neural" + 0.049\*"social" + 0.046\*"convolut" + 0.039\*"set" + 0.028\*"partial" + 0.027\*"group" + 0.020\*"use" + 0.018\*"recurr" + 0.018\*"observ"')

主要的主题为图像识别、知识图谱和神经网络。

在2013年-2017年，团队成员变为：

['Shiguang Shan', 'Xilin Chen', 'Ruiping Wang ', 'Zhiwu Huang']

论文的主题为：

{

12 : 1.47497,

24 : 1.00624,

17 : 0.865135,

21 : 0.857064,

1 : 0.54943,

……

}

前三个主题为：

(12,

'0.278\*"learn" + 0.073\*"deep" + 0.041\*"represent" + 0.040\*"activ" + 0.029\*"random" + 0.027\*"process" + 0.027\*"task" + 0.024\*"multi" + 0.023\*"transfer" + 0.022\*"model"')

(24,

'0.144\*"video" + 0.103\*"onlin" + 0.071\*"retriev" + 0.040\*"hash" + 0.039\*"depth" + 0.033\*"cross" + 0.029\*"modal" + 0.024\*"binari" + 0.017\*"dataset" + 0.015\*"analysi"')

(17,

'0.135\*"recognit" + 0.082\*"3d" + 0.068\*"action" + 0.061\*"face" + 0.054\*"improv" + 0.040\*"use" + 0.038\*"space" + 0.025\*"train" + 0.025\*"prior" + 0.022\*"model"')

主要涉及的主题是深度学习、视频处理和图像识别。

可以看到这个团队在2007-2012和2013-2017两个阶段，团队成员更换了两名，但是主题变化不大，主要是深度学习和图像识别。

## 总结

本次作业运用了频繁模式挖掘和隐含狄利克雷模型对DBLP数据库进行分析，在工作前期需要对数据进行预处理、分析等准备，工作中期需要根据作业要求，基于频繁项集、主题模型等来设计“团队”等复杂抽象的模型，工作后期需要对得到的结果进行分析，得出结论。可以看到，“团队”模型可以较好地展现团队成员之间的关系，并可以基于时间段对团队成员和研究主题的变迁进行刻画。