RankClus 的实现及在 DBLP 数据集上的实验

陈梓立<wander4096@gmail.com> 1500012726

概要

本次实验参考孙艺洲，韩家炜所著的《异构信息网络挖掘：原理和方法》一书中对 RankClus 算法的描述，实现了用于在异构网络中基于排名的聚类算法 RankClus，并在 DBLP 数据集上针对 2011 到 2018 年的 article 类资料，基于其 journal 属性(期刊，作为异构网络的属性 1)和 author 属性(作者，作为异构网络的属性 2)实践 RankClus 算法

数据提取

数据提取逻辑在代码文件 extractor.py 中，使用 LibXML2 的包装库，分析 dblp.xml 的内容，从中提取出 2011 年到 2018 年的 article 类型的资料

在此之后，根据作者发表文章的数量计算出权威值，最终生成的异构网络中只包含权威值为前 20000 的作者

数据提取方面的心得体会包括熟悉了 XML 文件的提取，尤其是与 DTD 文件配合的解析. dblp.xml 中有大量 HTML 转移字符，需要结合 DTD 文件才能正常解析

下载了 dblp.xml 及其 DTD 文件后，数据提取通过 `python3 <extractor> <xml> <dtd> <output>` 执行

RankClus 的概要

RankClus 算法的实践逻辑在代码文件 rankclus.py 中，主要包括数据读取，网络构建，按照算法迭代计算和最后的结果输出

// TODO

在项目根目录下执行 `python3 code/rankclus.py data/extract.csv` 运行算法，各项参数如迭代次数可通过修改 rankclus.py 头部的全局常量配置来改变

RankClus 的各部分算法

权威排名算法

算法包括两部分，排名和聚类，其共同基础都是排名分数的计算，根据参考资料实现了所谓的权威排名算法

这里主要从概念上说明这个函数的原理和内容，我们已有的信息是异构网络中刊物和作者的联系信息以及作者合著的联系信息，再此基础上我们基于三条简单的规则来设计一个更能利用正反馈作用的排名函数

1. 排名靠前的作者在排名靠前的刊物发表了许多的论文
2. 排名靠前的刊物吸引许多排名靠前的作者发表论文
3. 一个作者如果有许多高排名的合作者，他的排名因此被提高

综合考虑以上三点得到代码注释中的公式，进行若干次迭代的权威排名计算，以期排名分数收敛

```
                      alpha*sum(journal_rank * journal2author) + (1-alpha)*sum(author_rank*author2author)
Formula : author_rank = -------------------------------------------------------------------------------------
                                                  sum(all author_rank)
```

```
                        sum(author_rank * journal2author)
Formula : journal_rank = -----------------------------------
                              sum(all journal_rank)
```

EM 算法

EM 算法的目标在于估计每个聚类 k 的大小比例 p_k

```
                normalize(sum(journal2authour * p_k * authour_rank * journal_rank))
Formula : p_k = ----------------------------------------------------------------------
                                 sum_journal2authour
```

重聚类算法

根据上一步得到的 k 的大小，考虑贝叶斯条件，期刊 x_i 分到聚类 k 的概率加和为 1，这样我们就可以根据贝叶斯算法求出一个正比例关系，最终求出期刊 x_i 分到聚类 k 的概率

```
                             journal_rank * p_k
Formular : pi_k_journal = ------------------
                              sum(pi_k_journal)
```

这样，每个 x_i 可表为 (pi_i_1, pi_i_2, ..., pi_i_k)，紧接着就可以应用最平凡的聚类算法，重新计算聚类中心并修改作者的聚类标签

```
    """
    Calculate cluster center
    """
    center = { k : np.zeros(K) for k in range(K) }
    for k in range(K):
        for journal in clusters[k]:
            center[k] += pi_k_journal[journal]
        center[k] /= len(clusters[k])

    """
    Generate new clusters
    """
    clusters = defaultdict(list)
    for journal in pi_k_journal:
```

```
        similarity = {
            i : np.sum(pi_k_journal[journal] * center[i]) / (
                np.sqrt(np.sum(pi_k_journal[journal] *
pi_k_journal[journal])) *
                np.sqrt(np.sum(center[i] * center[i]))
            )
            for i in range(K)
        }
        clusters[max(similarity, key=(lambda x :
similarity[x]))].append(journal)
```

RankClus 的实现优化

1. 读取数据的方式经过几轮的实验最终采用了 `journal;authour[;authour]` 的格式，这是为了利用 Python 的 CSV 模块快速的切分原数据；在此调整下最终的读取数据和建立网络的时间为 2 秒左右，而一开始采用的 `journal$authour[;authour]` 虽然读取代码更有表达力，但是由于需要在读取后人工 split 来 parse，总建网约 2 - 5 分钟
2. 聚合操作的优化．由于数值计算过程中涉及到丰富的矩阵运算，因此尽可能的采用了 numpy 的向量计算函数．但是计算过程并不一定是规整的向量，实际上只在 cluster 向量时有规整的 K 维向量，此外的矩阵大量元素为空，采用类 C 写法再对简单的聚合操作使用 `sum` 和 `any` 等函数提升表达能力．总之，对于复杂的聚合操作，在多次走读代码的过程中进行了尽可能多的优化，权衡了效率上和表达力上的收益(奇技淫巧估计用了不少，主要是用得多了，也没感觉特别奇怪)
3. 并行优化．在 Authority Rank 一步，可以同时计算多个 Cluster 的权威排名向量，因此使用了 multiprocessing 模块进行并行计算．在最后的结果输出中也要计算权威排名，也进行了并行化．此外，构建网络实际上也可以并行化，但是一共就 2 秒且一开始未考虑，因此决定不做多余的优化．同样，其他地方的代码不是非常值得并行化，或者并行化的指标不好确定．唯一可能可做的是 EM 算法的迭代中对 k 进行并行，目前 EM 步骤大约耗时 35 秒，并行后有望降低至 10 秒左右

RankClus 的实测性能

随机性

算法的聚类初始化时随机的，因此在主迭代次数 MT 较小时算法出现显著的不确定性，但是 MT 较大时算法理论上将收敛到一个合理的范围内，不过由于聚类初始化和各个向量的初始化时随机的，因此算法的确存在随机性，两次运行的结果不一定一致

各部分性能报告

对于某次运行的输出结果，构建网络耗时约 3 秒

主迭代循环中，每轮 Authority Rank 耗时约 10 秒，EM 耗时约 35 秒，Clustering 耗时可不计

测试数据共 436817 行，19.7 M，总体运行时间约 22 分钟

结果分析

由于测试的范围与论文和书上不一致，没有找到好的对比对象，只靠人工对比确定算法在一定程度上的正确性，例如像能把 `VLDB J.` 和 `PVLDB` 分到同一个聚类，高产的论文作者例如 `JiaWei Han` 能正确的排到高排名

附录：一次运行的输出结果

```
Compose Network : 2.997405
Ranking Turn 0
Enter Authority Rank : 3.019870
Enter EM : 12.685903
Enter Clustering : 49.181277
Ranking Turn 1
Enter Authority Rank : 49.513723
Enter EM : 58.867085
Enter Clustering : 95.318905
Ranking Turn 2
Enter Authority Rank : 95.652656
Enter EM : 105.076496
Enter Clustering : 141.617709
Ranking Turn 3
Enter Authority Rank : 141.949557
Enter EM : 151.268375
Enter Clustering : 186.847755
Ranking Turn 4
Enter Authority Rank : 187.175372
Enter EM : 196.391857
Enter Clustering : 232.029454
Ranking Turn 5
Enter Authority Rank : 232.361089
Enter EM : 241.589593
Enter Clustering : 278.075249
Ranking Turn 6
Enter Authority Rank : 278.422610
Enter EM : 287.876591
Enter Clustering : 323.515293
Ranking Turn 7
Enter Authority Rank : 323.844313
Enter EM : 333.300490
Enter Clustering : 369.009743
Ranking Turn 8
Enter Authority Rank : 369.336668
Enter EM : 378.586696
Enter Clustering : 414.626375
Ranking Turn 9
Enter Authority Rank : 414.964719
Enter EM : 424.361403
Enter Clustering : 460.675054
Ranking Turn 10
```

```
Enter Authority Rank : 461.010703
Enter EM : 470.437984
Enter Clustering : 506.505864
Ranking Turn 11
Enter Authority Rank : 506.831902
Enter EM : 516.072677
Enter Clustering : 551.525253
Ranking Turn 12
Enter Authority Rank : 551.853751
Enter EM : 560.356076
Enter Clustering : 595.864595
Ranking Turn 13
Enter Authority Rank : 596.201700
Enter EM : 605.550482
Enter Clustering : 641.795729
Ranking Turn 14
Enter Authority Rank : 642.138541
Enter EM : 651.542974
Enter Clustering : 687.070272
Ranking Turn 15
Enter Authority Rank : 687.407860
Enter EM : 696.753591
Enter Clustering : 732.157443
Ranking Turn 16
Enter Authority Rank : 732.482757
Enter EM : 741.783523
Enter Clustering : 777.191358
Ranking Turn 17
Enter Authority Rank : 777.534438
Enter EM : 786.916339
Enter Clustering : 823.040715
Ranking Turn 18
Enter Authority Rank : 823.370965
Enter EM : 832.794512
Enter Clustering : 869.029111
Ranking Turn 19
Enter Authority Rank : 869.359963
Enter EM : 878.896278
Enter Clustering : 914.794153
Ranking Turn 20
Enter Authority Rank : 915.130293
Enter EM : 924.671366
Enter Clustering : 959.992035
Ranking Turn 21
Enter Authority Rank : 960.320584
Enter EM : 969.672809
Enter Clustering : 1005.376749
Ranking Turn 22
Enter Authority Rank : 1005.707741
```

```
Enter EM : 1015.060074
Enter Clustering : 1050.418304
Ranking Turn 23
Enter Authority Rank : 1050.751138
Enter EM : 1060.125719
Enter Clustering : 1095.926710
Ranking Turn 24
Enter Authority Rank : 1096.263383
Enter EM : 1105.723786
Enter Clustering : 1141.597562
Ranking Turn 25
Enter Authority Rank : 1141.932403
Enter EM : 1151.326966
Enter Clustering : 1187.294931
Ranking Turn 26
Enter Authority Rank : 1187.635884
Enter EM : 1196.127076
Enter Clustering : 1231.554687
Ranking Turn 27
Enter Authority Rank : 1231.882939
Enter EM : 1240.341101
Enter Clustering : 1275.694404
Ranking Turn 28
Enter Authority Rank : 1276.031411
Enter EM : 1285.386801
Enter Clustering : 1320.747747
Ranking Turn 29
Enter Authority Rank : 1321.071216
Enter EM : 1330.413438
Enter Clustering : 1365.942338
---------- OUTPUT : 1366.276855 ----------
Cluster 0

With Journals:
  Applied Mathematics and Computation
  CoRR
  Neurocomputing
  J. Computational Applied Mathematics
  Computers & Mathematics with Applications
  Appl. Math. Lett.
  J. Applied Mathematics
  Sensors
  Bioinformatics
  Inf. Sci.

With Authors:
  Stevo Stevic
  Ju H. Park
  Ioannis K. Argyros
```

```
  Juan R. Torregrosa
  Alicia Cordero
  H. M. Srivastava
  Beny Neta
  Shouming Zhong
  Ángel Alberto Magreñán
  Ravi P. Agarwal

Cluster 1

With Journals:
  CoRR
  IEEE Trans. Wireless Communications
  IEEE Trans. Information Theory
  IEEE Trans. Signal Processing
  IEEE Trans. Communications
  IEEE Trans. Vehicular Technology
  IEEE Communications Letters
  IEEE Journal on Selected Areas in Communications
  IEEE Trans. Image Processing
  IEEE Access

With Authors:
  H. Vincent Poor
  Yoshua Bengio
  Bernhard Rumpe
  Robert Schober
  Mérouane Debbah
  Rodrigo C. de Lamare
  Petar Popovski
  Loet Leydesdorff
  Chunhua Shen
  Walid Saad

Cluster 2

With Journals:
  CoRR
  Neurocomputing
  Inf. Sci.
  Sensors
  IEEE Trans. Cybernetics
  IEEE Trans. Neural Netw. Learning Syst.
  IEEE Trans. Image Processing
  IEEE Trans. Geoscience and Remote Sensing
  IEEE Trans. Industrial Electronics
  Pattern Recognition

With Authors:
```

```
  Jinde Cao
  Huaguang Zhang
  Tingwen Huang
  Xuelong Li
  Licheng Jiao
  Peng Shi 0001
  Yurong Liu
  Xinbo Gao
  Witold Pedrycz
  Fuad E. Alsaadi

Cluster 3

With Journals:
  J. Complexity
  CoRR
  Applied Mathematics and Computation
  Math. Comput.
  Journal of Approximation Theory
  J. Computational Applied Mathematics
  Numerical Algorithms
  SIAM J. Numerical Analysis
  IACR Cryptology ePrint Archive
  Numerische Mathematik

With Authors:
  Henryk Wozniakowski
  Erich Novak
  Grzegorz W. Wasilkowski
  Aicke Hinrichs
  Ian H. Sloan
  Friedrich Pillichshammer
  Ioannis K. Argyros
  Josef Dick
  Shun Zhang
  Frances Y. Kuo

Cluster 4

With Journals:
  CoRR
  IEEE Trans. Image Processing
  Neurocomputing
  IEEE Trans. Multimedia
  IEEE Trans. Circuits Syst. Video Techn.
  Pattern Recognition
  ACM Trans. Graph.
  IEEE Trans. Pattern Anal. Mach. Intell.
  IEEE Trans. Cybernetics
```

```
  IEEE Trans. Neural Netw. Learning Syst.

With Authors:
  Dacheng Tao
  Shuicheng Yan
  Weisi Lin
  Xuelong Li
  Wen Gao 0001
  Qi Tian
  Ling Shao 0001
  Meng Wang 0001
  Lei Zhang 0006
  Xinbo Gao

Cluster 5

With Journals:
  CoRR
  IEEE Trans. Wireless Communications
  IEEE Trans. Vehicular Technology
  IEEE Trans. Communications
  IEEE Communications Letters
  IEEE Access
  IEEE Journal on Selected Areas in Communications
  IEEE Communications Magazine
  IEEE Trans. Signal Processing
  IET Communications

With Authors:
  Lajos Hanzo
  Mohamed-Slim Alouini
  Robert Schober
  Zhu Han
  Victor C. M. Leung
  H. Vincent Poor
  Zhiguo Ding
  Robert W. Heath Jr.
  George K. Karagiannidis
  Inkyu Lee

Cluster 6

With Journals:
  J. Comput. Physics
  CoRR
  SIAM J. Scientific Computing
  J. Sci. Comput.
  SIAM J. Numerical Analysis
  J. Computational Applied Mathematics
```

```
  Computers & Mathematics with Applications
  Applied Mathematics and Computation
  Sensors
  Math. Comput.

With Authors:
  George E. Karniadakis
  Chi-Wang Shu
  Jan Nordström
  Nikolaus A. Adams
  Michael Dumbser
  Dinshaw S. Balsara
  Frédéric Gibou
  Xiangyu Hu 0002
  Mikhail J. Shashkov
  Jianxian Qiu

Cluster 7

With Journals:
  IEICE Transactions
  CoRR
  IEEE Communications Letters
  IEEE Trans. Vehicular Technology
  Sensors
  IEEE Access
  Neurocomputing
  IEICE Electronic Express
  IACR Cryptology ePrint Archive
  Wireless Personal Communications

With Authors:
  Akira Matsuzawa
  Satoshi Goto
  Hyoung-Kyu Song
  Fumiyuki Adachi
  Masayuki Murata
  Kiyomichi Araki
  Masahiko Yoshimoto
  Kunihiro Asada
  Jiro Hirokawa
  Makoto Ando

Cluster 8

With Journals:
  CoRR
  Theor. Comput. Sci.
  Discrete Applied Mathematics
```

```
  Discrete Mathematics
  Algorithmica
  Graphs and Combinatorics
  J. Comb. Optim.
  Electronic Notes in Discrete Mathematics
  Journal of Graph Theory
  SIAM J. Discrete Math.

With Authors:
  Dieter Rautenbach
  Michael A. Henning
  Daniël Paulusma
  Petr A. Golovach
  Hong-Jian Lai
  Martin Milanic
  Jayme Luiz Szwarcfiter
  Zsolt Tuza
  Pinar Heggernes
  Dimitrios M. Thilikos

Cluster 9

With Journals:
  NeuroImage
  CoRR
  IEEE Trans. Med. Imaging
  Medical Image Analysis
  IEEE Trans. Biomed. Engineering
  PLoS Computational Biology
  Front. Neuroinform.
  Brain Connectivity
  J. Cognitive Neuroscience
  Neurocomputing

With Authors:
  Karl J. Friston
  Simon B. Eickhoff
  Vince D. Calhoun
  Paul M. Thompson
  Dinggang Shen
  Bruce Fischl
  Peter T. Fox
  Arthur W. Toga
  Stephen M. Smith
  Thomas E. Nichols

Cluster 10

With Journals:
```

```
  CoRR
  Wireless Personal Communications
  IEEE Access
  Sensors
  IEEE Trans. Vehicular Technology
  IJDSN
  IEEE Communications Letters
  IEICE Transactions
  Security and Communication Networks
  J. Network and Computer Applications

With Authors:
  Ramjee Prasad
  Gwo-Jiun Horng
  Fathi E. Abd El-Samie
  Vidhyacharan Bhaskar
  Borhanuddin Mohd Ali
  Muhammad Khurram Khan
  Maode Ma
  Jong Hyuk Park
  Hyung Yun Kong
  Alagan Anpalagan

Cluster 11

With Journals:
  IACR Cryptology ePrint Archive
  CoRR
  J. Cryptology
  Electronic Colloquium on Computational Complexity (ECCC)
  IEEE Trans. Information Theory
  Des. Codes Cryptography
  IEICE Transactions
  IEEE Trans. Information Forensics and Security
  J. Cryptographic Engineering
  Security and Communication Networks

With Authors:
  Amit Sahai
  Brent Waters
  Vinod Vaikuntanathan
  Rafail Ostrovsky
  Debdeep Mukhopadhyay
  François-Xavier Standaert
  Daniel J. Bernstein
  Sanjam Garg
  Daniel Wichs
  Ivan Damgård
```

```
Cluster 12

With Journals:
  CoRR
  IEEE Trans. Parallel Distrib. Syst.
  IEEE Trans. Computers
  IEEE Trans. VLSI Syst.
  IEEE Trans. on CAD of Integrated Circuits and Systems
  IEEE Trans. Mob. Comput.
  Future Generation Comp. Syst.
  IEEE/ACM Trans. Netw.
  IEEE Trans. Vehicular Technology
  IEEE Trans. on Circuits and Systems

With Authors:
  Albert Y. Zomaya
  Yunhao Liu
  Haiying Shen
  Jie Wu 0001
  Irith Pomeranz
  Song Guo
  Hai Jin
  Krishnendu Chakrabarty
  Jiannong Cao
  Xiang-Yang Li 0001

Cluster 13

With Journals:
  CoRR
  Automatica
  IEEE Trans. Automat. Contr.
  Systems & Control Letters
  IEEE Trans. Contr. Sys. Techn.
  IEEE Trans. Industrial Electronics
  Int. J. Control
  Neurocomputing
  IEEE Trans. Cybernetics
  Int. J. Systems Science

With Authors:
  Miroslav Krstic
  Karl Henrik Johansson
  Lihua Xie
  Peng Shi 0001
  Ian R. Petersen
  Dragan Nesic
  Andrew R. Teel
  Frank Allgöwer
```

```
  Ling Shi
  Huijun Gao

Cluster 14

With Journals:
  CoRR
  PVLDB
  IEEE Trans. Knowl. Data Eng.
  VLDB J.
  RFC
  Neurocomputing
  Inf. Sci.
  World Wide Web
  Knowl. Inf. Syst.
  IEEE Trans. Image Processing

With Authors:
  Lei Chen 0002
  Jeffrey Xu Yu
  Xuemin Lin
  Philip S. Yu
  Christian S. Jensen
  Guoliang Li 0001
  Jiawei Han 0001
  Gang Chen 0001
  Xiaofang Zhou
  Divesh Srivastava
```