**互联网数据挖掘第一次作业**

——基于ANN论文数据集计算paper/auther/venue的Pagerank

学 号：

姓 名：

时 间：2017年10月15日

1. **paper的pagerank**
   1. 使用数据

aanrelease2014\aan\release\2014\network\paper-citation-network.txt

aanrelease2014\aan\release\2014\ paper\_ids.txt

* 1. 步骤
     + 1. 将paper\_id转化为整数id，即paper\_ids.txt中的行数为paper的数值id，输出文件pid.txt；
       2. 将paper-citation-network.txt中的字符id转化为数字id，输出文件pcn.txt，步骤1、2由paper2id.cpp实现；
       3. 将pcn.txt中的引用关系读入到引用关系矩阵，应用pagerank算法：**r = PTr**，迭代计算pagerank向量r。利用快排算法排序后，输出paper\_pagerank.txt文件。

补充说明：由于总的paper数为24628，本人的计算机内存有限，无法生成double[24628][24628]，所以使用了BitSet数据结构，即每篇paper按位来表示应用关系，引用则为1，未引用则为0。此时还需要记录paper的引出数目，以便矩阵取值时给出分数值。

* 1. 结果

前十名如下，根据paper\_ids.txt可以查到paper名称：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Paper\_id | pagerank | 被引用 | 引用 |
| 1 | A88-1019 | 0.00991224 | 237 | 1 |
| 2 | A88-1030 | 0.00850365 | 5 | 2 |
| 3 | C86-1033 | 0.00574866 | 15 | 0 |
| 4 | J90-2002 | 0.00485857 | 201 | 1 |
| 5 | P86-1022 | 0.00364922 | 4 | 0 |
| 6 | J93-2004 | 0.00334808 | 1008 | 8 |
| 7 | J93-2003 | 0.00326693 | 783 | 8 |
| 8 | J86-3001 | 0.00244021 | 360 | 6 |
| 9 | P83-1019 | 0.00182947 | 37 | 3 |
| 10 | C88-1016 | 0.00176549 | 30 | 1 |

分析：以上结果是在迭代误差小于10-8时得出。前十名中被引用数最多的是J93-2004，但它引用了第一名A88-1019，大大增加了第一名的权威性。第二名A88-1030的被引用数只有五，但因为它获得了排名最高的A88-1019唯一的引用，所以也能排名靠前。其余文章或因为被引用数高，或因为被排名高的文章引用，所以能名列此榜。

1. **author的pagerank**
   1. 使用数据

aanrelease2014\aan\release\2014\author-citation-network.txt

aanrelease2014\aan\release\2014\ author\_ids.txt

* 1. 步骤

由于author\_ids.txt中已经给出了author的数值id，且author-citation-network.txt是数值对数值的引用，故直接将其导入引用关系矩阵，使用迭代算法计算pagerank值。

* 1. 前十位结果如下：

|  |  |  |
| --- | --- | --- |
| 排名 | 作者 | Pagerank值 |
| 1 | Church,Kenneth Ward | 0.000896999 |
| 2 | Manning,Christopher D. | 0.000733203 |
| 3 | Brill,Eric | 0.000719151 |
| 4 | Della Pietra,Vincent | 0.000698859 |
| 5 | Marcus,Mitchell P. | 0.000679526 |
| 6 | Della Pietra,Stephen A. | 0.000609113 |
| 7 | Pereira,Fernando | 0.000587391 |
| 8 | Mercer,Robert L. | 0.000568196 |
| 9 | Brown,Peter F. | 0.000514456 |
| 10 | Klein,Dan | 0.000513995 |

1. **venue的pagerank**
   1. 使用数据

aanrelease2014\aan\release\2014\acl-metadata.txt

aanrelease2014\aan\release\2014\network\paper-citation-network.txt

* 1. 步骤
     + 1. 给venue设置id，统计acl-metadata.txt中出现的所有venue名字，用c++中set模板收集，再遍历set，输出文件venue\_ids.txt;
       2. 建立paper到venue的映射。根据acl-metadata.txt中paper和venue的对应关系，输出文件paper2venue.txt；
       3. 建立paper到venue\_id的映射。根据venue\_ids.txt中venue和venue\_id的对应关系，读入文件paper2venue.txt，输出paper2venue\_id.txt；
       4. 根据paper2venue\_id.txt，将paper-citation-network.txt中paper对paper的引用转化为venue到venue的引用，输出文件vcn.txt；
       5. 利用vcn.txt中的引用关系，建立引用关系矩阵，迭代计算出pagerank值。
  2. 结果

|  |  |  |
| --- | --- | --- |
| 排名 | 会议 | Pagerank值 |
| 1 | ACL | 0.192136 |
| 2 | COLING | 0.116393 |
| 3 | CL | 0.106772 |
| 4 | EMNLP | 0.0643265 |
| 5 | NAACL | 0.0478692 |
| 6 | HLT | 0.0370799 |
| 7 | CoNLL | 0.0329681 |
| 8 | ANLP | 0.0268565 |
| 9 | EACL | 0.025399 |
| 10 | Workshop On Speech And Natural Language | 0.0219211 |