数据结构实验二文档

文档索引及搜索查询

2011013251 软件 11 吕婉琪

目录

实验目的

给定 2000 个规范化网页(暂 定搜狐博客的 2000 个网页),应用实验 1 的接口建立倒排文档,并对 倒排文档中的词典用 B 树进行组织,形成二层索引结构。

实验环境

Windows 7 + Microsoft Visual Studio 2010

抽象数据类型说明

class CharString

```
long length;                                  //字符串长度
elemType* data;                               //字符内容

long index(CharString &cs);                   //模式匹配
CharString subString(long pos, long len);     //截断字符串
void concat(CharString &cs);                  //连接字符串
```

class BTree

```
BTNode* root;                                                  //根结点

void init();                                                   //初始化
bool search(CharString term,int &num, BTNode* &p);             //搜索
void insert(Keyword* k, int num, BTNode* &p);                  //插入
void edit(BTNode* &p, int num, int Doc_ID);                    //修改
void deletetree(BTNode* p);                                    //删除 B-树
```

class DocLink

```
int termID;                                  //单词的 ID
DocNode* head;                               //头结点

void init();                                 //初始化
bool search(int DocID, DocNode * &p);        //搜索
void insert(int DocID);                      //插入
void edit(int times, DocNode* p);            //修改
```

实验基本算法和流程

首先对 2000 个网页用实验一接口进行分词，将结果存储到“分词.txt”中。然后根据所有分词结果建立倒排文档，对词典构建一棵 4 叉 B-树，形成二级索引结构，并利用其进行查找。

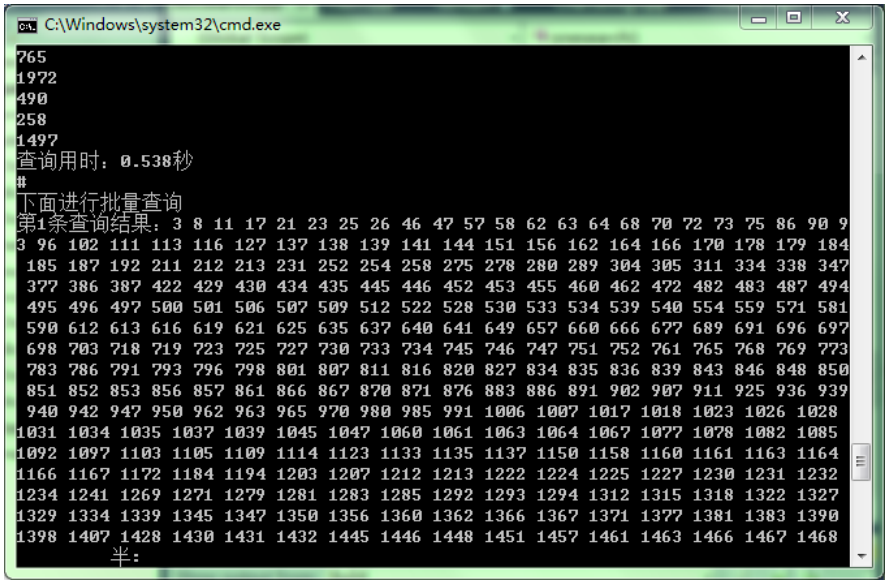
另用 hash 函数进行查找，hash 函数为按位取余构造的，解决冲突的方法为顺序查找。

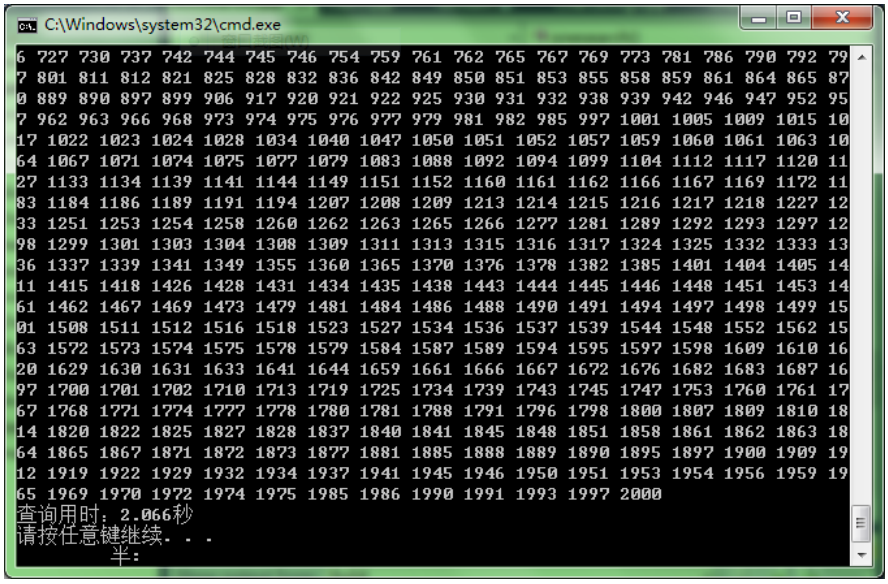
程序输入输出及操作说明

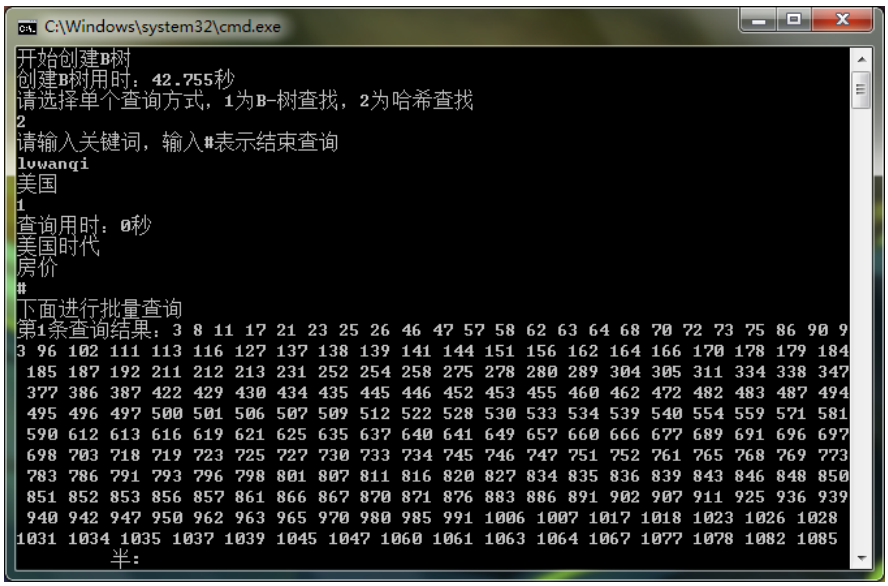
程序的输入为“分词.txt”以及“查询输入.txt”，输出为“分词 B 树.txt”以及“查询结果.txt”。

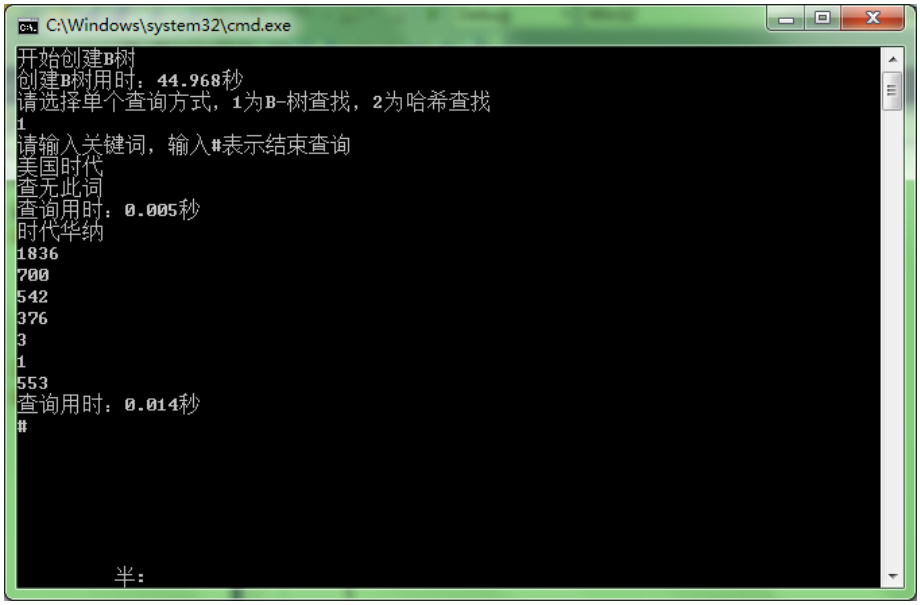
实验测试结果

具体结果见 src 文件夹中的各 txt。以下是部分截图。


C:\Windows\system32\cmd.exe
765
1972
490
258
1497
查询用时：0.538秒
#
下面进行批量查询
第1条查询结果：3 8 11 17 21 23 25 26 46 47 57 58 62 63 64 68 70 72 73 75 86 90 9
3 96 102 111 113 116 127 137 138 139 141 144 151 156 162 164 166 170 178 179 184
185 187 192 211 212 213 231 252 254 258 275 278 280 289 304 305 311 334 338 347
377 386 387 422 429 430 434 435 445 446 452 453 455 460 462 472 482 483 487 494
495 496 497 500 501 506 507 509 512 522 528 530 533 534 539 540 554 559 571 581
590 612 613 616 619 621 625 635 637 640 641 649 657 660 666 677 689 691 696 697
698 703 718 719 723 725 727 730 733 734 745 746 747 751 752 761 765 768 769 773
783 786 791 793 796 798 801 807 811 816 820 827 834 835 836 839 843 846 848 850
851 852 853 856 857 861 866 867 870 871 876 883 886 891 902 907 911 925 936 939
940 942 947 950 962 963 965 970 980 985 991 1006 1007 1017 1018 1023 1026 1028
1031 1034 1035 1037 1039 1045 1047 1060 1061 1063 1064 1067 1077 1078 1082 1085
1092 1097 1103 1105 1109 1114 1123 1133 1135 1137 1150 1158 1160 1161 1163 1164
1166 1167 1172 1184 1194 1203 1207 1212 1213 1222 1224 1225 1227 1230 1231 1232
1234 1241 1269 1271 1279 1281 1283 1285 1292 1293 1294 1312 1315 1318 1322 1327
1329 1334 1339 1345 1347 1350 1356 1360 1362 1366 1367 1371 1377 1381 1383 1390
1398 1407 1428 1430 1431 1432 1445 1446 1448 1451 1457 1461 1463 1466 1467 1468
半:


C:\Windows\system32\cmd.exe
6 727 730 737 742 744 745 746 754 759 761 762 765 767 769 773 781 786 790 792 79
7 801 811 812 821 825 828 832 836 842 849 850 851 853 855 858 859 861 864 865 87
0 889 890 897 899 906 917 920 921 922 925 930 931 932 938 939 942 946 947 952 95
7 962 963 966 968 973 974 975 976 977 979 981 982 985 997 1001 1005 1009 1015 10
17 1022 1023 1024 1028 1034 1040 1047 1050 1051 1052 1057 1059 1060 1061 1063 10
64 1067 1071 1074 1075 1077 1079 1083 1088 1092 1094 1099 1104 1112 1117 1120 11
27 1133 1134 1139 1141 1144 1149 1151 1152 1160 1161 1162 1166 1167 1169 1172 11
83 1184 1186 1189 1191 1194 1207 1208 1209 1213 1214 1215 1216 1217 1218 1227 12
33 1251 1253 1254 1258 1260 1262 1263 1265 1266 1277 1281 1289 1292 1293 1297 12
98 1299 1301 1303 1304 1308 1309 1311 1313 1315 1316 1317 1324 1325 1332 1333 13
36 1337 1339 1341 1349 1355 1360 1365 1370 1376 1378 1382 1385 1401 1404 1405 14
11 1415 1418 1426 1428 1431 1434 1435 1438 1443 1444 1445 1446 1448 1451 1453 14
61 1462 1467 1469 1473 1479 1481 1484 1486 1488 1490 1491 1494 1497 1498 1499 15
01 1508 1511 1512 1516 1518 1523 1527 1534 1536 1537 1539 1544 1548 1552 1562 15
63 1572 1573 1574 1575 1578 1579 1584 1587 1589 1594 1595 1597 1598 1609 1610 16
20 1629 1630 1631 1633 1641 1644 1659 1661 1666 1667 1672 1676 1682 1683 1687 16
97 1700 1701 1702 1710 1713 1719 1725 1734 1739 1743 1745 1747 1753 1760 1761 17
67 1768 1771 1774 1777 1778 1780 1781 1788 1791 1796 1798 1800 1807 1809 1810 18
14 1820 1822 1825 1827 1828 1837 1840 1841 1845 1848 1851 1858 1861 1862 1863 18
64 1865 1867 1871 1872 1873 1877 1881 1885 1888 1889 1890 1895 1897 1900 1909 19
12 1919 1922 1929 1932 1934 1937 1941 1945 1946 1950 1951 1953 1954 1956 1959 19
65 1969 1970 1972 1974 1975 1985 1986 1990 1991 1993 1997 2000
查询用时：2.066秒
请按任意键继续. . .
半:


C:\Windows\system32\cmd.exe
开始创建B树
创建B树用时：42.755秒
请选择单个查询方式，1为B-树查找，2为哈希查找
2
请输入关键词，输入#表示结束查询
lvwanqi
美国
1
查询用时：0秒
美国时代
房价
#
下面进行批量查询
第1条查询结果：3 8 11 17 21 23 25 26 46 47 57 58 62 63 64 68 70 72 73 75 86 90 9
3 96 102 111 113 116 127 137 138 139 141 144 151 156 162 164 166 170 178 179 184
185 187 192 211 212 213 231 252 254 258 275 278 280 289 304 305 311 334 338 347
377 386 387 422 429 430 434 435 445 446 452 453 455 460 462 472 482 483 487 494
495 496 497 500 501 506 507 509 512 522 528 530 533 534 539 540 554 559 571 581
590 612 613 616 619 621 625 635 637 640 641 649 657 660 666 677 689 691 696 697
698 703 718 719 723 725 727 730 733 734 745 746 747 751 752 761 765 768 769 773
783 786 791 793 796 798 801 807 811 816 820 827 834 835 836 839 843 846 848 850
851 852 853 856 857 861 866 867 870 871 876 883 886 891 902 907 911 925 936 939
940 942 947 950 962 963 965 970 980 985 991 1006 1007 1017 1018 1023 1026 1028
1031 1034 1035 1037 1039 1045 1047 1060 1061 1063 1064 1067 1077 1078 1082 1085
半:


C:\Windows\system32\cmd.exe
开始创建B树
创建B树用时：44.968秒
请选择单个查询方式，1为B-树查找，2为哈希查找
1
请输入关键词，输入#表示结束查询
美国时代
查无此词
查询用时：0.005秒
时代华纳
1836
700
542
376
3
1
553
查询用时：0.014秒
#
半：


查询结果.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
第3条查询结果：
3 4 9 16 24 25 27 30 45 54 61 62 66 68 69 76 78 82 85 86 87 90
106 107 109 115 124 126 127 129 133 134 138 139 147 149 156 167
170 177 179 185 186 191 197 198 209 224 228 231 238 244 248 254
268 273 275 290 301 302 305 306 312 324 328 333 339 344 350 354
359 362 364 366 368 372 375 376 379 382 384 385 386 395 398 403
404 412 419 420 422 424 429 430 431 432 436 440 451 452 467 470
471 472 478 481 483 484 499 507 508 510 512 513 514 520 526 527
529 545 548 556 558 563 566 577 582 587 590 591 604 611 612 616
618 619 625 628 629 630 631 632 633 636 637 639 654 656 664 665
668 675 676 678 681 684 687 698 707 710 715 718 726 727 730 737
742 744 745 746 754 759 761 762 765 767 769 773 781 786 790 792
797 801 811 812 821 825 828 832 836 842 849 850 851 853 855 858
859 861 864 865 870 889 890 897 899 906 917 920 921 922 925 930
931 932 938 939 942 946 947 952 957 962 963 966 968 973 974 975
976 977 979 981 982 985 997 1001 1005 1009 1015 1017 1022 1023
1024 1028 1034 1040 1047 1050 1051 1052 1057 1059 1060 1061 1063
1064 1067 1071 1074 1075 1077 1079 1083 1088 1092 1094 1099 1104
1112 1117 1120 1127 1133 1134 1139 1141 1144 1149 1151 1152 1160
1161 1162 1166 1167 1169 1172 1183 1184 1186 1189 1191 1194 1207
1208 1209 1213 1214 1215 1216 1217 1218 1227 1233 1251 1253 1254
1258 1260 1262 1263 1265 1266 1277 1281 1289 1292 1293 1297 1298
1299 1301 1303 1304 1308 1309 1311 1313 1315 1316 1317 1324 1325

分数申请

大项	小项	分数（%）	申请分数（%）
数据结构	词典	20%	20%
	倒排文档	15%	15%
功能	建立倒排文档	15%	15%
	效率	10%	9%
	正确性	20%	20%
文档		15%	14%
加分		15%	12%

总申请分数：105

（加分项目：用 hash 实现查找）