超酷算法-BK树

前几天无意间遇到一个博客，觉得写得挺好的，自己之前的时候有个不好的习惯，那就是遇到了好资源第一反应就是收藏起来然后却很少再看！！这是坏习惯，要改！于是今天就开始通读了，读的第二篇是BK树。觉得有点意思，于是乎就萌发了写个博客啥的，但是呢，我发现已经有人翻译了。那还干嘛重复发明轮子呢，鉴于原作者声明禁止转载，那就算了吧，想看原文的来这里。

下面简单说明一下这个算法，确实不难，只是思路有点巧妙。

BK树解决一个什么问题呢，简单而言就是找相似字符串，比如说"book"跟"boon"是不是只差一个字母，很相似了吧。

我们先来定义相似：采用编辑距离来度量两个字符串之间的相似程度。字符串A和字符串B的编辑距离就是至少需要几次操作（删除一个字母，插入一个字母，更换一个字母）才能使得A变成B。上面提到的"book"以及"boon"的编辑距离就是1，因为只需要将字母'k'更新为'n'就可以达到目的了。

接下来我们来看编辑距离的一个性质，我们用L(A,B)来表示字符串A和字符串B之间的编辑距离。那么我们为了找到与A距离不超过m的字符串C，那么它与字符串B的距离为多少呢？答案是L(A,B)-m <= L(B,C) <= L(A,B)+m。为什么？m步之内A、C可以相互转换，而L(B,C)步之内B、C可以相互转换，于是乎m+L(B,C)步之内，A、B之间必然可以转换，于是有L(A,B) <= L(B,C)+m；同理可知 L(B,C) <= L(A,B)+m。

那么这样一来的话，BK树就可以出场了。BK树的边是有编号的，编号值就是边的两个节点直接的编辑距离。

我们先在字符串集合中任选一个字符串Z作为根节点，然后每次从集合中取出一个字符串X，将其插入树中。插入规则是这样的，首先计算X与根节点Z的编辑距离L(X,Z)，然后将这个节点插入到Z的编号为L(X,Z)的孩子那边；递归直到到达X可以成为叶子节点。

我们查找字符串A的相似字符串的时候（假设编辑距离为2以内就算相似），那么从根节点开始寻找，先计算L(Z,A)，这个时候我们就知道了与A编辑距离为2的字符串只可能存在于Z的编号为L(Z,A)-2到编号为L(Z,A)+2之间的那些子树里面，于是乎就递归查找去吧。

苯苯吹雪
关注 - 9
粉丝 - 29
+加关注

1 0

« 上一篇：HDFS error
» 下一篇：局部性原理的点滴应用场景 use of localityprinciple

posted on 2014-12-04 22:49 苯苯吹雪 阅读(671) 评论(8) 编辑 收藏

评论

#1楼 2014-12-24 15:12 netxiaosheng

学长，你上博士了？

支持(0) 反对(0)

#2楼[楼主] 2014-12-25 13:26 苯苯吹雪

公告

昵称：苯苯吹雪
园龄：5年
粉丝：29
关注：9

导航

博客园
首页
新随笔
联系
订阅XML
管理

@ netxiaosheng
您是？

支持(0) 反对(0)

#3楼 2014-12-26 14:48 netxiaosheng

@ 苯苯吹雪

> 引用
> @netxiaosheng您是？

哦。一个大三的学生，搜算法的时候看到你的博客的，我以为你在HIT，上研究生，看错了。不知道如何入手学习算法。

支持(0) 反对(0)

#4楼[楼主] 2014-12-26 23:18 苯苯吹雪

@ netxiaosheng
哈哈，你没看错，我是HIT的准研究生。我的经验是照着骆老师的课件和算法导论（这个有在线视频公开课的）自己一步步玩，之后就应该可以称之为入门了。

支持(0) 反对(0)

#5楼 2014-12-28 07:32 netxiaosheng

@ 苯苯吹雪

> 引用
> @netxiaosheng哈哈，你没看错，我是HIT的准研究生。我的经验是照着骆老师的课件和算法导论（这个有在线视频公开课的）自己一步步玩，之后就应该可以称之为入门了。

哦。是算法导论国外的那个视频，和骆老师的课件一起看的对吗？骆老师没有视频吧？我打算明年考哈工大的，不知道学长是研究生几年级了，可不可以给与一些经验，让我少走一些弯路，我在大庆东油，离得很近，去过几次工大，想考那里。那我更要关注你的博客了，千万不要不跟新哈，要不然就找不到学长了。

支持(0) 反对(0)

#6楼[楼主] 2014-12-31 12:08 苯苯吹雪

@ netxiaosheng
视频是指MIT的教学视频。然后，考研的话，欢迎过来我工，但是我没有考研不太清楚具体细节，抱歉啦

支持(0) 反对(0)

#7楼 2015-01-03 12:33 netxiaosheng

@ 苯苯吹雪
嗯

支持(0) 反对(0)

#8楼 2015-01-21 08:50 beaglebone

@苯苯吹雪 你好，你的这个算法看懂了，就是利用三角形法则，与A距离为2的节点为B，A与Z节点距离为L(A,Z)，那么B与Z之间的距离L(B,Z)满足：
L(Z,A)-2<=L(Z,B)<=L(Z,A)+2
但是把X插入Z树中，为什么需要进行递归，使得X成为Z的叶子节点。

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

OPENCV(1)
unisim(1)
xilinx(1)
历史(1)
更多

随笔档案

2015年4月 (1)
2015年2月 (1)
2015年1月 (2)
2014年12月 (10)
2014年11月 (2)
2014年8月 (5)
2014年7月 (1)
2014年6月 (3)
2014年5月 (4)
2014年4月 (9)
2014年3月 (3)
2014年2月 (1)
2013年12月 (7)
2013年11月 (4)
2013年10月 (22)
2013年9月 (3)
2013年7月 (1)
2013年6月 (2)
2013年4月 (4)
2013年3月 (12)
2013年2月 (2)
2012年12月 (10)
2012年11月 (5)

最新评论

1. Re:BM算法 Boyer-Moore高量实现代码详解与算法详解

badchar规则解释严重有误。参见

--Ra

2. Re:算术编码Arithmetic Coding高质量代码实现详解

楼主，你好，我试了一下你的代码以运行，但是有个问题不太明白。我入了30个字符'0'，然后我将code数的内容打印了出来，数了一下，co中非0x00的字符已经不止30个了哪这样的话根本就没.....

--layerlearn

3. Re:25匹马中找出跑的最快的三马

解释的很清晰

--beaglebo

4. Re:BM算法 Boyer-Moore高量实现代码详解与算法详解

谢谢，在坏字符配图错误时，给出了释说明。

--萧

5. Re:BM算法 Boyer-Moore高量实现代码详解与算法详解

9 for (i = 0; i < m - 1; ++i)10 bmBc[x[i]] = m - i - 1;12 13 } for (i = 0; i <= m -.....

--lantul

最新IT新闻:
· 软银入股Uber"双管齐下" Uber估值究竟几何？
· 网易第三季度营收19亿美元 同比增长35.5%
· 特斯拉麻烦不断 马斯克发飙：空头小人想整死我们
· 医药电商风口已失：堆钱买流水，成一单亏200元
· 3.92元/股 基金经理第三次下调乐视网估值
» 更多新闻...

阅读排行榜

1. BM算法　Boyer-Moore高质实现代码详解与算法详解(9203)
2. WIN8 小米刷机 启动qcCoInstaller.dll是出现错误(835
3. 机器学习（一）：梯度下降、神网络、BP神经网络(6969)
4. 整数划分 Integer Partition（-(4732)
5. 算术编码Arithmetic Coding - 质量代码实现详解(4524)

评论排行榜

1. 后缀树系列二:线性时间内构建后树（包含代码实现）(10)
2. 超酷算法-BK树(8)
3. 整数划分 Integer Partition（-(5)
4. BM算法　Boyer-Moore高质实现代码详解与算法详解(4)
5. 机器学习（一）：梯度下降、神网络、BP神经网络(2)

推荐排行榜

1. BM算法　Boyer-Moore高质实现代码详解与算法详解(5)
2. LCS最长公共子序列（最优线性间O(n)）(4)
3. 整数划分 Integer Partition（-(2)
4. 算术编码Arithmetic Coding - 质量代码实现详解(1)
5. String Reduction(1)