超酷算法（1）：BK树

2014/10/22 · 开发 · 1 评论 · 算法

本文由 伯乐在线 - 威士忌 翻译。未经许可，禁止转载！
英文出处：notdot。欢迎加入翻译组。

这是『超酷算法』系列的第一篇文章。基本上，任何一种算法我觉得都很酷，尤其是那些不那么明显简单的算法。

BK树或者称为Burkhard-Keller树，是一种基于树的数据结构，被设计于快速查找近似字符串匹配，比方说拼写检查器，或模糊查找，当搜索”aeek”时能返回”seek”和”peek”。为何BK-Trees这么酷，因为除了穷举搜索，没有其他显而易见的解决方法，并且它能以简单和优雅的方法大幅度提升搜索速度。

BK树在1973年由Burkhard和Keller第一次提出，论文在这《Some approaches to best match file searching》。这是网上唯一的ACM存档，需要订阅。更细节的内容，可以阅读这篇论文《Fast Approximate String Matching in a Dictionary》。

在定义BK树之前，我们需要预先定义一些操作。为了索引和搜索字典，我们需要一种比较字符串的方法。编辑距离（ Levenshtein Distance）是一种标准的方法，它用来表示经过插入、删除和替换操作从一个字符串转换到另外一个字符串的最小操作步数。其它字符串函数也同样可接受（比如将调换作为原子操作），只要能满足以下一些条件。

本条件：

- d(x,y) = 0 <-> x = y (假如x与y的距离为0，则x=y)
- d(x,y) = d(y,x) (x到y的距离等同于y到x的距离)
- d(x,y) + d(y,z) >= d(x,z)

上述条件中的最后一条被叫做三角不等式（Triangle Inequality）。三角不等式表明x到z的路径不可能长于另一个中间点的任何路径（从x到y再到z）。看下三角形，你不可能从一点到另外一点的两侧再画出一条比它更短的边来。

编辑距离符合基于以上三条所构造的度量空间。请注意，有其它更为普遍的空间，比如欧几里得空间（Euclidian Space），编辑距离不是欧几里得的。既然我们了解了编辑距离（或者其它类似的字符串距离函数）所表达的度量的空间，再来看下Burkhard和Keller所观察到的关键结论。

假设现在我们有两个参数，query表示我们搜索的字符串，n表示字符串最大距离，我们可以拿任意字符串test来跟query进行比较。调用距离函数得到距离d，因为我们知道三角不等式是成立的，所以所有结果与test的距离最大为d+n，最小为d-n。

由此，BK树的构造就相当简单：每个节点有任意个子节点，每条边有个值表示编辑距离。所有子节点到父节点的边上标注n表示编辑距离恰好为n。比如，我们有棵树父节点是”book”和两个子节点”rook”和”nooks”，”book”到”rook”的边标号1，”book”到”nooks”的边上标号2。

从字典里构造好树后，取任意单词作为树的根节点。无论何时你想插入新单词时，计算该单词与根节点的编辑距离，并且查找数值为d(neweord, root)的边。递归得与各子节点进行比较，直到没有子节点，你就可以创建新的子节点并将新单词保存在那。比如，插入”boon”到刚才上述例子的树中，我们先检查根节点，查找d(“book”，“boon”) = 1的边，然后检查标号为1的边的子节点，得到单词”rook”。我们再计算距离d(“rook”，“boon”)=2，则将新单词插在”rook”之后，边标号为2。

在树中做查询，计算单词与根节点的编辑距离d，然后递归查找每个子节点标号为d-n到d+n（包含）的边。假如被检查的节点与搜索单词的距离d小于n，则返回该节点并继续查询。

BK树是多路查找树，并且是不规则的（但通常是平衡的）。试验表明，1个查询的搜索距离不会超过树的5-8%，并且2个错误查询的搜索距离不会超过树的17-25%，这可比检查每个节点改进了一大步啊！需要注意的是，如果要进行精确查找，也可以非常有效地通过简单地将n设置为0进行。

回顾这篇文章，写的有点长哈，似乎比我预期中的要复杂。希望你在阅读之后，也能感受到BK树的优雅和简单。

关于作者：威士忌

前算法爱好者，现游戏从业者。（新浪微博：@棍子WisKey 个人网站）

相关文章

- BFPRT 算法（TOP-K问题） · 2
- 决策树算法及实现
- 漫画算法：什么是 B 树？ · 2
- 漫画算法：什么是跳跃表？ · 5
- 七大查找算法

可能感兴趣的话题

- 有同做 Android for ROS 的小伙...
- 程序员清晰的职业规划会有多长？ · 4
- layout布局优化
- GreenDao多表联查
- PHP 的可能 · 1
- 怎么提高组织语言能力和表达能力？求指导... · 1

登录后评论　新用户注册

直接登录

最新评论

鼎郦Eming　2015/01/24

翻译的水平不是很好哦，有待提高 加油

赞　回复

- 本周热门文章
- 本月热门文章
- 热门标签

0 开发者死后，他的开源项目会有人继续...

1 不懂技术的管理者，给你们扫盲软件开...

2 10 个鲜为人知的 Linux 命令（5）

3 30 个实例详解 TOP 命令

4 10 个鲜为人知的 Linux 命令（3）

5 10 个鲜为人知的 Linux 命令（4）

6 分布式事务的一种实现方式--状态流转

7 QA 请勿忘初心

8 读懂 MySQL 执行计划

9 2017 最优秀的十大 Linux 服务器...

业界热点资讯 **更多 »**

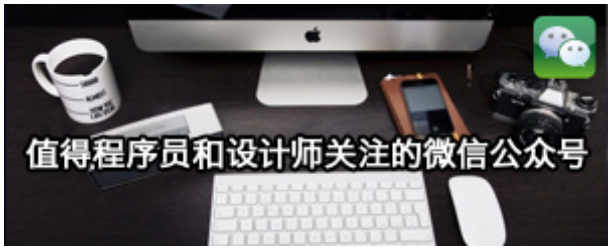
1 天前 · 14

[慕尼黑放弃 Linux，2020 年或将全面迁入 Windows]
19 小时前 · 3

[N 次跳票后，Fedora 27 正式版终于发布了]
20 小时前 · 2

[最新的 Java SE 平台和 JDK 版本发布计划]
1 天前 · 3

[TIOBE 11 月编程语言排行榜，脚本语言怎么了？]
1 天前 · 4



精选工具资源 **更多资源 »**

[Whitewidow：SQL 漏洞自动扫描工具]
[数据库] · 2

机器学习

静态代码分析工具清单：公司篇
静态代码分析

HotswapAgent：支持无限次重定义运行时类与资源
开发流程增强工具

静态代码分析工具清单：开源篇（各语言）
静态代码分析

关于伯乐在线博客

在这个信息爆炸的时代，人们已然被大量、快速并且简短的信息所包围。然而，我们相信：过多“快餐”式的阅读只会令人“虚胖”，缺乏实质的内涵。伯乐在线内容团队正试图以我们微薄的力量，把优秀的原创文章和译文分享给读者，为“快餐”添加一些“营养”元素。

快速链接
网站使用指南 »
问题反馈与求助 »
加入我们 »
网站积分规则 »
网站声望规则 »

关注我们

新浪微博：@伯乐在线官方微博
RSS：订阅地址
推荐微信号

合作联系
Email：bd@Jobbole.com
QQ： 2302462408（加好友请注明来意）

小组 – 好的话题、有启发的回复、值得信赖的圈子
头条 – 分享和发现有价值的内容与观点
相亲 – 为IT单身男女服务的征婚传播平台
资源 – 优秀的工具资源导航
翻译 – 翻译传播优秀的外文文章
文章 – 国内外的精选文章
设计 – UI,网页，交互和用户体验
iOS – 专注iOS技术分享
安卓 – 专注Android技术分享
前端 – JavaScript, HTML5, CSS
Java – 专注Java技术分享
Python – 专注Python技术分享