中山大學
SUN YAT-SEN UNIVERSITY

本科生毕业论文

题　　目：基于压缩文本的字符串近似查询算法研究

院　　系：数据科学与计算机学院

专　　业：计算机科学与技术

学生姓名：李贝瑀

学　　号：11318030

指导教师：林瀚（讲师）

二〇一六年四月

表一　毕业论文（设计）开题报告

论文 (设计) 题目: 基于压缩文本的字符串近似查询算法研究

(简述选题的目的、思路、方法、相关支持条件及进度安排等)

选题目的

如何高效的存储大量字符串以及快速检索其中相近的子集，在基因组序列、Wiki 条目等问题中起着关键的作用。目前较多的方法只侧重于其中一点，即牺牲查找速度实现压缩存储，或建立大规模的索引加快搜索。因此，找出一种兼备存储效率及检索速度的方法很有实用意义。

思路及方法

对于文本压缩，比较常见的方法是对于某些近似的字符串，选出一个字符串作为它们的参考。该方法的问题在于，对于需要检索的字符串，只能找出与其近似的参考，而不是近似的字符串子集。在这个方法的基础上加以改进，我们可以构建分层结构的多重参考，即每个字符串存在多个参考，参考间存在层次化的依赖关系。对于检索，需要先找到匹配的参考，通过对参考中预处理信息的考察，得出所有与询问串编辑距离小于 k 的字符串集合。问题的难点在于，找出合适的算法与数据结构，用于构造字符串间的参考关系，在保证较高压缩率的同时，兼顾检索的速度。

进度安排

2015 年 12 月查看文献，总结他人在该类问题上的思路与技巧

2016 年 03 月尝试多种算法及数据结构，测试比较并完成论文撰写

学生签名:　　年　　月　　日

指导教师意见:

海量数据的存储和检索是数据科学的重要问题，作者的选题有学术价值，也有技术难度。

1、同意开题 ( ) 2、修改后开题 ( ) 3、重新开题 ( )

指导教师签名:　　年　　月　　日

表二　毕业论文（设计）过程检查情况记录表

指导教师分阶段检查论文的进展情况（要求过程检查记录不少于 3 次）：

第 1 次检查

学生总结：

阅读了几篇关于字符串压缩、字符串 k 近似匹配的论文，学习了解决这两个问题较为常见的思路、方法。

指导教师意见：

第 2 次检查

学生总结：

大致构建了压缩的算法、查询的算法框架，考虑用后缀数组或者后缀树做字符串匹配。

从部分结果进行扩展时，似乎可以使用函数式线段树查询区间，但空间复杂度略高。

指导教师意见：

第 3 次检查

学生总结：

最终选择了较为简洁的后缀自动机作为主要的数据结构进行字符串匹配，区间的查询可用区间树维护。

完成了总体的代码，压缩率一般，继续考虑较为细节的策略。

指导教师意见：

总体完成情况	指导教师意见：

	1、按计划完成，完成情况优（） 2、按计划完成，完成情况良（） 3、基本按计划完成，完成情况合格（） 4、完成情况不合格（） 指导教师签名:　　年　　月　　日

表三　毕业论文 (设计) 答辩情况登记表

答辩人	李贝瑀	专业	计算机科学与技术
论文（设计）题目	基于压缩文本的字符串近似查询算法研究		
答辩小组成员			
答辩记录:			
记录人签名:　　年　　月　　日			

学术诚信声明

本人所呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。本毕业论文的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

本人签名:　　日期:

[论文摘要]

对于海量数据的存储与检索，是数据科学领域的重要问题，在构建人类基因库、Wiki文库历史版本维护等问题中起着关键作用。近年来，越来越多的研究开始将两者结合起来，在保证数据压缩存储的同时，兼备检索的能力。本文旨在利用一种较为简洁的数据结构——后缀自动机，对于可添加的字符串集合，动态构建包含多个引用的层次化索引结构，其中：1）引用的选择由程序自身决策，而非人工选择；2）支持基于编辑距离的近似字符串匹配。经证实，对于索引结构的字符串压缩方案来说，寻找空间最优解是一个 NP-hard 问题。因此，本文的最终目标在于尽量优化存储空间，在满足较高压缩率的同时，兼顾检索的速度。

[关键词] 数据压缩；近似匹配；索引；后缀自动机

[**Abstract**]

Storing and searching of massive data is an important issue in the field of data science, which plays a key role in problems as construction of the human gene pool and in maintenance problems about historical versions in Wiki library. Recently, more and more researchers have been beginning to combine them. In other words, when it comes to the compressed storage of data, the searching ability is also demanded. This paper aims to use a relatively compact data structure - suffix automaton, to dynamically build the hierarchical index structure containing multiple reference, for the set of strings that can expand, in which: 1) the reference is selected by its own decision-making procedures, rather than artificial selection; 2) it is supported to match approximate string based on their edit-distance. It was confirmed that for the index structure string compression scheme, finding the optimal solution in space is an NP-hard problem. Therefore, the ultimate goal of this paper is trying to optimize storage space to meet the higher compression ratio while taking a efficent searching method into account.

[**Key Words**] Compression; approximate matching; indexing; suffix automaton

目录

第一章　引言

在信息科学领域中，字符串及相关问题一直以来都是研究的热点，其中最具代表性的当属匹配 [19] 与压缩 [] 问题。在数据量越来越大的今天，直接对字符串本身进行存储将占用大量空间，如何在其中检索有用的信息也是一大难题。那么对于庞大的字符串集合，要如何将其压缩至尽量小的空间，却又可以尽量快的进行检索呢？

1.1　背景

近年来，随着信息技术在基因工程方面的普及，越来越多的 DNA 序列问题依赖计算机来处理。例如在研究基因突变时，由于突变率较低，必须记录下大量实验数据，再找出其中与原 DNA 序列近似的片段。其他方面，网络知识库的加速更迭，使得历史版本的存储带来挑战。在 Wiki 条目中，对于某个名词，可能存在许多类似的别称，需要一种近似检索的策略。

在字符串检索领域中有一种较为重要的方式——k 近似匹配：在一个字符串集合中，找出与模式串之间编辑距离不超过 k 的所有子串，其中编辑距离表示将模式串通过插入、修改、删除变为目标串的最小修改次数。传统的 k 近似匹配基本只用到两种方法：1）对于集合内的字符串分别处理；2）将集合内的字符串拼接起来处理。这两种方法的局限性在于，存储所占用的空间至少是所有字符串长度之和。由于自然语言的特性，许多语句之间具有极强的关联性，可以通过其中的某一句话的片段，表达其他句子的部分信息，从而达到压缩存储的目的。因此，越来越多研究开始尝试先选择一些字符串作为索引，再将与之相似的串通过引用进行压缩的方法。此时，k 近似匹配变成了两个问题：1）在索引集合中检索；2）找出与该索引串相似的引用串。这种方法的问题在于，只有索引中的匹配是精确的，引用串只是与索引串相似，而非模式串。

因此，找出一种将两者结合起来，在保证数据压缩存储的同时，兼备高效检索能力的算法是十分有意义的。

1.2　相关成果

我们首先简单介绍一下前人有关字符串压缩算法的研究成果，以及这一技术在解决各领域有关检索问题时的应用。

1.2.1 字符串压缩

1.3 本文工作

本文结合了前人的研究结果，尝试构建动态的多引用可搜索索引模型，支持动态扩展的字符串集合，通过多引用提高压缩率，自动决策引用的选择以及提供高效的搜索策略。构建思路的难点在于：1）每次要在原有结构的基础上进行简单修改即可完成更新，需要适合的数据结构；2）每个引用串可以引用多个不同的索引，需要在字符串集合中进行快速匹配，设计尽量优化空间的选择策略；3）需要快速的 k 近似匹配算法，以及合适的数据结构维护引用的关系，从而递推出全部结果；4）特殊情况的处理。下面我们将对上述问题依次讨论。

第二章　模版使用说明

2.1　正常使用

本模版主要组成部分有:

- `main.tex` 为主文件，毕业论文的内容应放在这个文件中。
- `main.bib` 为参考文献记录。
- `sysuthesis.bst` 为参考文献样式文件。
- `sysuthesis.cls` 为文类文件，应与主文件放在相同的目录中。
- `image/logo.png` 为用在封面的校徽。

主文件 `main.tex` 基本上是自解释的。在正常使用时，只用按照源文件中注释在正确位置填写各种基本信息、开题报告内容和中英文摘要内容，然后把注释 “现在开始正文” 到注释 “篇末注” 之间部分替换为正文，再按照注释在正确位置加入参考文献、致谢和附录等。最后用 `xelatex` 编译即可:

```
xelatex main
bibtex main
xelatex main
xelatex main
```

。当然也可以用 `makefile` 或用脚本简化编译过程。

2.2　配置

在使用本模版前，请先保证已安装以下宏包:

- `ctex`
- `tocloft`
- `calc`
- `graphicx`
- `amsmath`
- `amssymb`
- `amsthm`
- `listings`
- `subfig`

- longtable
- endnotes
- algorithm2e
- hyperref
- placeins

为了让 algorithm2e 宏包支持中文，请把该宏包的文件 algorithm2e.sty (在我的系统中，它在 /usr/share/texlive/texmf-dist/tex/latex/algorithm2e/) 替换为本模版附带的同名文件。不需要描述算法时，可以在 sysuthesis.cls 中去掉以下代码

```
\RequirePackage[chinese,onelanguage,noline,noend,
  linesnumbered]{algorithm2e}
```

并在主文件中去掉以下代码

```
\listofalgorithms
```

(源文件有注释提示)。

如果在 windows 编译且希望使用微软的字体时，请把 sysuthesis.cls 中以下代码

```
\LoadClass[adobefonts,a4paper,openright,cs4size,fancyhdr
  ]{ctexbook}[2010/01/22]
```

改为

```
\LoadClass[winfonts,a4paper,openright,cs4size,fancyhdr]{
  ctexbook}[2010/01/22]
```

。

如果想使用开源字体 Fandol (假定已安装它)，请把 ctex 宏包的文件 ctex-xecjk-adobefonts.def (在我的系统中，它在 /usr/share/ texlive/texmf-dist/tex/l 替换为本模版附带的同名文件。

关于参考文献的 bib 条目格式参考 main.bib [1][2][3][4][5][6][7][8][9][10][11][12]。

第三章　写作时的注意事项

本章列出一些本模板未能规范也不太适合以注释形形式写在源文件的要求，请使用者注意。

3.1　毕业论文的撰写内容与要求

正文一般包括以下几个方面:

1. 引言或背景

   引言是论文正文的开端, 应包括毕业论文选题的背景、目的和意义; 对国内外研究现状和相关领域中已有的研究成果的简要评述; 介绍本项研究工作研究设想、研究方法或实验设计、理论依据或实验基础; 涉及范围和预期结果等。要求言简意赅, 注意不要与摘要雷同或成为摘要的注解。

2. 主体

   论文主体是毕业论文的主要部分, 必须言之成理, 论据可靠, 严格遵循本学科国际通行的学术规范。在写作上要注意结构合理、层次分明、重点突出, 章节标题、公式图表符号必须规范统一。论文主体的内容根据不同学科有不同的特点, 一般应包括以下几个方面:

   - 毕业论文 (设计) 总体方案或选题的论证;
   - 毕业论文 (设计) 各部分的设计实现, 包括实验数据的获取、数据可行性及有效性的处理与分析、各部分的设计计算等;
   - 对研究内容及成果的客观阐述, 包括理论依据、创新见解、创造性成果及其改进与实际应用价值等;
   - 论文主体的所有数据必须真实可靠, 凡引用他人观点、方案、资料、数据等, 无论曾否发表, 无论是纸质或电子版, 均应详加注释。自然科学论文应推理正确、结论清晰; 人文和社会学科的论文应把握论点正确、论证充分、论据可靠, 恰当运用系统分析和比较研究的方法进行模型或方案设计, 注重实证研究和案例分析, 根据分析结果提出建议和改进措施等。

3. 结论

   结论是毕业论文的总结, 是整篇论文的归宿, 应精炼、准确、完整。结论应着重阐述自己的创造性成果及其在本研究领域中的意义、作用, 还可进一

步提出需要讨论的问题和建议。

3.2 毕业论文的撰写格式要求

除有特殊要求的专业外, 毕业论文正文一般不少于 5000 字。各专业可根据需要确定具体的文字和字数要求, 并报教务处备案。

正文各部分的标题应简明扼要, 不使用标点符号。

名词术语的要求:

- 科学技术名词术语尽量采用全国自然科学名词审定委员会公布的规范词或国家标准、部标准中规定的名称, 尚未统一规定或叫法有争议的名词术语, 可采用惯用的名称。
- 特定含义的名词术语或新名词、以及使用外文缩写代替某一名词术语时, 首次出现时应在括号内注明其含义, 如: OECD (Organisation for Economic Co-operation and Development) 代替经济合作发展组织。
- 外国人名一般采用英文原名, 可不译成中文, 英文人名按姓前名后的原则书写, 如:CRAY P, 不可将外国人姓名中的名部分漏写, 例如: 不能只写 CRAY, 应写成 CRAY P。一般很熟知的外国人名 (如牛顿、爱因斯坦、达尔文、马克思等) 可按通常标准译法写译名。

物理量名称、符号与计量单位的要求:

- 论文中某一物理量的名称和符号应统一, 一律采用国务院发布的《中华人民共和国法定计量单位》。单位名称和符号的书写方式, 应采用国际通用符号。
- 在不涉及具体数据表达时允许使用中文计量单位如“千克”。
- 表达时刻应采用中文计量单位, 如“下午 3 点 10 分”, 不能写成“3h10min”, 在表格中可以用“3:10PM”表示。
- 物理量符号、物理量常量、变量符号用斜体, 计量单位符号均用正体。

数字的要求:

- 无特别约定情况下, 一般均采用印度——阿拉伯数字表示。
- 年份一律使用 4 位数字表示。
- 小数的表示方法: 一般情形下, 小于 1 的数, 需在小数点之前加 0。但当某些特殊数字不可能大于 1 时 (如相关系数、比率、概率值), 小数点之前的 0 要去掉, 如 r=.26,p<.05。
- 统计符号的格式: 一般除 、 、 、 、 以及 V 等符号外, 其余统计符号一律以斜体字呈现, 如 ANCOVA , ANOVA , MANOVA , N , nl , M , SD , F , p , r 等。

公式的要求:

- 公式应另起一行写在稿纸中央。一行写不完的长公式, 最好在等号处转行,

如做不到这一点, 可在运算符号 (如“+”、“-”号) 处转行, 等号或运算符号应在转行后的行首。

- 公式的编号用圆括号括起, 放在公式右边行末, 在公式和编号之间不加虚线。公式可按全文统编序号, 也可按章独立序号, 如 (49) 或 (4.11)。采用哪一种序号应和图序、表序编法一致。不应出现某章里的公式编序号, 有的则不编序号。子公式可不编序号, 需要引用时可加编 a、b、c....., 重复引用的公式不得另编新序号。公式序号必须连续, 不得重复或跳缺。
- 文中引用某一公式时, 写成“由式 (16.20)”。

表格的要求 (表3.1为一个也许不合格的例子):

- 表格必须与论文叙述有直接联系, 不得出现与论文叙述脱节的表格。表格中的内容在技术上不得与正文矛盾。
- 每个表格都应有自己的标题和序号。标题应写在表格上方正中, 不加标点, 序号写在标题左方。
- 全文的表格可以统一编序, 也可以逐章单独编序。采用哪一种方式应和插图、公式的编序方式统一。表序必须连续, 不得跳缺。
- 表格允许下页接写, 接写时标题省略, 表头应重复书写, 并在右上方写“续表 ××”。多项大表可以分割成块, 多页书写, 接口处必须注明“接下页”、“接上页”、“接第 × 页”字样。
- 表格应放在离正文首次出现处最近的地方, 不应超前和过分拖后。

表 3.1: 各个倾斜校正方法性能对比

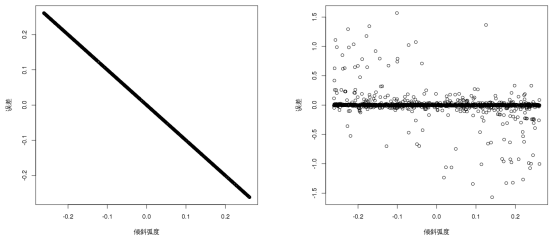
	成功返回率	平均误差 (弧度)	均方误差 (弧度)	误差中位数 (弧度)	运行时间中位数 (毫秒)
参照方法	**100.00**%	0.1298	0.1504	0.1286	**0**
分片填涂方法	93.48%	0.0621	0.1947	0.0061	50
分片覆盖方法	99.10%	0.0154	**0.0299**	0.0073	300
投影方法	99.35%	**0.0068**	0.0335	**0.0033**	197
交错数方法	99.35%	0.0101	0.0458	0.0035	198
霍夫变换方法	99.35%	0.1448	0.1827	0.0452	99
行间相关方法	93.81%	0.2115	0.3811	0.0346	1615
最近邻方法	96.71%	0.0883	0.1207	0.0690	67

图的要求 (图3.1为一个也许不合格的例子):

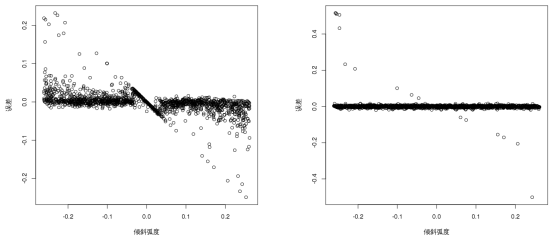
- 插图应与文字内容相符, 技术内容正确。所有制图应符合国家标准和专业标准。对无规定符号的图形应采用该行业的常用画法。
- 每幅插图应有标题和序号, 全文的插图可以统一编序, 也可以逐章单独编序, 如: 图 45 或图 6.8。采取哪一种方式应和表格、公式的编序方式统一。

图序必须连续, 不重复, 不跳缺。

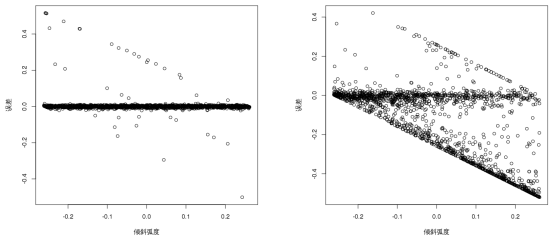
- 由若干分图组成的插图, 分图用 a、b、c..... 标序。分图的图名以及图中各种代号的意义, 以图注形式写在图题下方, 先写分图名, 另起行写代号的意义。
- 图与图标题、图序号为一个整体, 不得拆开排版为两页。当页空白不够排版该图整体时, 可将其后文字部分提前, 将图移至次页最前面。
- 对坐标轴必须进行文字标示, 有数字标注的坐标图必须注明坐标单位。



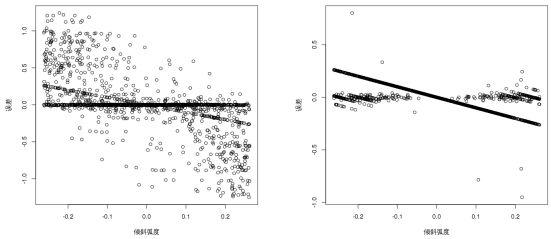
(a) 参照方法　　(b) 分片填涂方法



(c) 分片覆盖方法　　(d) 投影方法



(e) 交错数方法　　(f) 霍夫变换方法



(g) 行间相关方法　　(h) 最近邻方法

图 3.1: 各个倾斜校正方法的残差图

第四章　结论

4.1　取得成果

一个大概不为人知的模版，可以认为对于人类无任何贡献。

4.2　后续工作

此模版有以下已知局限性：

- 篇末注只支持最多 10 个
- 有时用户可能要直接修改 cls 文件
- 遗留了一些格式化代码在 tex 文件

参考文献：

[1] 作者 A, 作者 B, 作者 C, 等. 文章名 [J]. 期刊, 2000, **4**(3):12–87.

[2] 作者 A, 作者 B, 作者 C. 书名 [M]. 北京: 出版社, 1999.

[3] 作者. 标题 [D]. 合肥: 学校, 1999.

[4] 作者. 标题 [A]. 刊名 [C], 北京: 出版社, 1998:123–323.

[5] 作者. 标题 [D]. 合肥: 学校, 1999.

[6] 作者. 文献题目 [EB/OL]. `http://www.gnu.org/`, 2012-10-12.

[7] 报告者. 报告题目 [R]. 报告地: 报告会主办单位, 报告年份.

[8] 作者. 数据库 [DB/OL]. `http://www.gnu.org/`, 1999-10-3.

[9] 作者. 软件 [CP/DK]. `http://www.gnu.org/`, 2012-10-12.

[10] 代号. 标题 [S]. 出版地: 出版单位, 1999.

[11] 作者. 标题 [N]. 报纸, 2015-2-1.

[12] 作者 A, 作者 B. 标题:国家, 代号 [P]. 2001-12-2.

致谢

致谢内容

作者

2016 年 4 月 12 日

附录 A　记号约定

记号	说明
$A \cup B$	集合 A 与集合 B 的并集
$A \cap B$	集合 A 与集合 B 的交集
$A \setminus B$	集合 A 与集合 B 的差集
$A \times B$	集合 A 与集合 B 的直积
$A/\simeq$	集合 A 关于等价关系 $\simeq$ 的商集

毕业论文成绩评定记录

指导教师评语：   成绩评定： 指导教师签名：　　年　　月　　日
答辩小组或专业负责人意见：   成绩评定： 签名（章）：　　年　　月　　日
院系负责人意见：   成绩评定： 签名（章）：　　年　　月　　日