目录

[第一章 研发内容 1](#_Toc463447493)

[1 技术背景 1](#_Toc463447494)

[2 技术指标 2](#_Toc463447495)

[第二章 功能模块 2](#_Toc463447496)

[第三章 附录 2](#_Toc463447497)

[3.1倒排索引 2](#_Toc463447498)

# 第一章 研发内容

## 1 技术背景

常见文档信息搜索方式包括：网络搜索、企业搜索和P2P搜索。一个具体的应用场景为：用户向搜索引擎输入一个包含目标关键字的搜索请求并由后者以表单形式返回一个有序的文档列表。不同于数据库中的结构化查询[[1]](#footnote-1)，文档内容通常没有固定结构，在其上执行查找作业时需要按字符比较方式从中提取与关键字匹配的信息片断。文本索引能够提高文档信息搜索效率，被广泛运用于商业、学术和生物医疗领域。例如：谷歌[[2]](#endnote-1)和百度[[3]](#endnote-2)等常用搜索引擎以及BWA和SOAP等经典生物基因对齐软件分别采用倒排索引和后缀索引来提高搜索效率。

倒排索引和后缀索引是全文索引[[4]](#footnote-2)，两者均能够在O(nlogn+occ)的时间内返回指定关键字在文档中的所有出现。本项目采用全文索引构建一个用于日文翻译的垂直型信息检索系统，对给定的输入关键字、词、句，返回指定上下文中（词库、句库、原文、译文）的相关内容（解释、示例）。

## 2 技术指标

索引构造：构造速度、访问速度、索引大小、更新速度、可扩展

索引查找：单用户响应时间、多用户响应时间

# 第二章 功能模块

# 第三章 附录

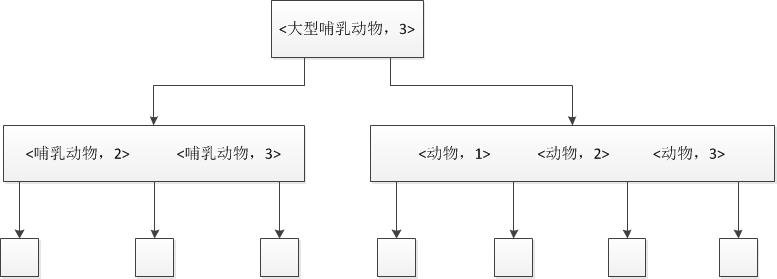
## 3.1倒排索引

倒排索引可将一个原子搜索项（例如：单词）映射到一个包含该项的被索引单元集合（例如：文件集）。可使用B树构建倒排索引，树中节点包含<word, location>并以此二段优先权排序。下面以表1数据为输入，构造并查找关键字“动物”：

|  |  |
| --- | --- |
| 表1 | |
| 文件集 | {这是一只动物、这是一只哺乳动物、这是一只大型哺乳动物} |
| 词袋 | {动物、哺乳动物、大型哺乳动物} |

索引构造过程包括以下3步：

1. 生成元组。组合无序元组序列，得到<动物，1>，<动物，2>，<动物，3>，<哺乳动物，2>，<哺乳动物，3>，<大型哺乳动物，3>。
2. 排序元组。依次按字典序和整数序对各元组的首字段和次字段排序，得到有序元组序列：<哺乳动物，2>，<哺乳动物，3>，<大型哺乳动物，3>，<动物，1>，<动物，2>，<动物，3>。
3. 构建B树。构建3阶B树如下。



索引查找过程包括以下2步：

1. 判定存在。从B树根节点出发，访问编号为1的节点，得到所有包含“动物”的三个元组。
2. 获取位置。从步骤1中查找得到的元组中取出包含关键字的文件编号，使用字符串匹配算法（例如：KMP）在指定文件中查找“动物”的出现。

1. 不包含数据库中支持的文本属性的精确/模糊查询。 [↑](#footnote-ref-1)
2. [] google search engine, http://www.google.com. [↑](#endnote-ref-1)
3. [] baidu search engine, https://www.baidu.com. [↑](#endnote-ref-2)
4. 严格地说，倒排索引不是全文索引。其假设输入关键字来自一个预定义词袋并对词袋做预处理。 [↑](#footnote-ref-2)