信息检索系统设计与实现报告

徐奕璋 2022211373

1.系统概述

本系统是一个基于Python实现的中文信息检索系统，支持对学术论文进行高效检索。系统包含两个核心模块：

- **索引构建模块 (create_re_table.py)**: 处理原始数据并构建优化的倒排索引
- **查询处理模块 (query.py)**: 实现自然语言查询处理和结果展示

系统特点

- 支持中文分词和停用词过滤
- 实现基于TF-IDF和位置加权的复合评分机制
- 提供搜索结果人工评价功能
- 支持搜索结果匹配显示

2.数据采集与存储

数据源

- 使用python爬虫获取北邮学报论文数据
- 数据规模：约5000篇文档
- 数据格式：JSON结构存储论文元数据

文档结构示例

```
{
  "title": "论文标题",
  "abstract": "摘要内容",
  "author": ["作者1", "作者2"],
  "keyword": ["关键词1", "关键词2"],
  "url": "原文链接",
  "date": "发表日期"
}
```

3.倒排索引构建技术实现

3.1 分词处理

- 使用jieba进行中文分词
- 加载cn_stopwords.txt停用词表进行过滤
- 空字段安全处理机制

```
def safe_segment(text):
    if not text.strip():
```

```
            return []
        return [t for t in jieba.cut(text) if t.strip() and t not in zh_stop]
```

3.2 位置信息记录

- 为每个词语记录在文档中的位置信息
- 支持标题、摘要、作者和关键词四个字段
- 位置信息用于后续相关性计算

```
# 标题位置记录示例
title_tokens = safe_segment(title)
tp = defaultdict(list)
for i, t in enumerate(title_tokens):
    tp[t].append(i)  # 记录词语在标题中的位置
```

3.3 倒排索引结构

```
{
  "关键词": {
    "文档ID": {
      "title_positions": [位置列表],
      "abstract_positions": [位置列表],
      "author_positions": [位置列表],
      "keyword_positions": [位置列表],
      "score": 权重分数
    }
  }
}
```

3.4 创新性评分机制

原始TF-IDF算法 传统TF-IDF（词频-逆文档频率）公式如下： $TFIDF = TF \times IDF$

其中：

- TF (词频)：词在文档中出现的频率 $TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
- IDF (逆文档频率)：衡量词在整个语料库中的重要性 $IDF(t,D) = \log\frac{N}{|{d \in D: t \in d}|}$

传统方法存在局限性：

1. 高频词主导结果
2. 对专有名词不够敏感
3. 分数范围不一致

本系统的优化公式 我们采用复合评分公式： $score = sigmoid\left(\frac{\log(1 + \sqrt{TF} \times IDF^3)}{3}\right)$

分步计算过程：

1. **TF计算优化：**

```
doc_tf = math.log(1 + term_count) / math.log(1 + doc_length)
tf = math.log(1 + doc_tf_sum)
```

   - 使用对数变换平滑词频，抑制高频词影响
   - 考虑文档长度归一化，解决长文档优势问题

2. **IDF增强：** $raw = log(1 + \sqrt{tf} * idf^3)$

   - 使用IDF的三次方（IDF^3）显著增强专有名词的权重
   - 对常见词（高DF）进行更强惩罚

3. **非线性变换:**

```
raw_score = math.log(1 + math.sqrt(tf) * idf * idf * idf)
```

   - $\sqrt{TF}$ 平衡词频影响
   - $\log(1+x)$ 压缩分数范围

4. **Sigmoid归一化:**

```
def sigmoid(x):
   return 2 / (1 + math.exp(-x)) - 1

final_score = sigmoid(raw_score / 3)
```

   - 将分数映射到(0,1)区间
   - 参数3控制曲线陡峭度，优化分数分布

优化原理分析

优化点	传统方法问题	本系统解决方案	效果
高频词主导	常见词得分过高	$\sqrt{TF}$ + 对数变换	抑制高频词影响
稀有词敏感度	稀有词区分度不足	IDF^3增强	显著提升稀有词权重
分数范围	无界且不一致	Sigmoid归一化	统一到(0,1)区间
文档长度偏差	长文档优势明显	长度归一化	公平对待不同长度文档

实际效果 在raw_scores.txt文件中记录的计算示例：

词	TF	IDF	Raw_Score	Final_Score

词	TF	IDF	Raw_Score	Final_Score
Kronecker	0.477671	6.891626	5.425916	0.718394
COPE	0.207245	7.807307	5.382859	0.714904
SNR	1.540260	5.613331	5.395954	0.715969
拥塞	2.501668	4.702499	5.108822	0.691837
方案	5.198349	2.137825	3.147450	0.481223
构造	3.769341	3.256142	4.219889	0.606468
解决	4.390652	2.432834	3.439516	0.517744

> 通过这种复合评分机制，系统在保持TF-IDF核心思想的同时，显著提升了稀有词的重要性，优化了相关性排序效果，为高质量检索结果奠定了基础。

4. 查询处理与检索模型

4.1 查询处理流程

1. 中文分词与停用词过滤
2. 识别作者名和关键词
3. 多字段联合检索
4. 相关性评分排序
5. 结果高亮展示

4.2 向量空间模型优化

- **字段权重差异化：**

```
TITLE_WEIGHT = 5.0
AUTHOR_WEIGHT = 50.0
KEYWORD_WEIGHT = 10.0
ABSTRACT_WEIGHT = 1.0
```

- **位置加权策略：**
  - 标题匹配 > 作者匹配 > 关键词匹配 > 摘要匹配
  - 作者和关键词匹配获得更高权重

4.3 结果展示优化

- **多字段高亮显示：**

```
def highlight_title(title, hit_list):
    title_terms = [term for (field, term) in hit_list if field == 'title']
    unique_terms = list(set(title_terms))
    unique_terms.sort(key=lambda x: len(x), reverse=True)
```

```
    tokens = list(jieba.cut(title))
    highlighted = [False] * len(tokens)
    for term in unique_terms:
        for i, token in enumerate(tokens):
            if token == term and not highlighted[i]:
                tokens[i] = f'【{token}】'
                highlighted[i] = True
    return ''.join(tokens)
```

- **结构化输出：**
  - 相关度分数
  - 高亮标题
  - 作者列表
  - 摘要片段
  - 关键词
  - URL链接
  - 发表日期

5. 系统评估方案

1. 用户输入rate命令触发评价
2. 展示当前搜索结果供评估
3. 收集多行文本反馈
4. 结构化存储评价日志：

```
[评价时间] 2025-05-29 11:51:05
搜索词：张伟

Top3 搜索结果：
1. 相关度：50.0

    标题：失序可控的实时多径传输负载分发模型

    作者：【张伟】 雷为民 李广野 关云冲 李浩

    摘要：提出了一种用于实时多径传输的失序可控的负载分发模型.根据目的端反馈的各条路径的传输质量以及路径之间的失序信息，在源端动态更新各条路径上的负载分配份额，在满足与路径传输质量成正比的负载均衡的同时，最小化目的端的数据包失序风险.仿真结果表明，负载分发模型可有效降低目的端由数据包失序导致的丢包率.

    关键词：多径传输 应用层中继 负载分发 数据包失序 实时传输

    URL：https://journal.bupt.edu.cn/CN/10.13190/j.jbupt.2016.03.018

    日期：2016-06-28


2. 相关度：50.0
```

```
    标题：社交网络用户身份关联及其分析

    作者：孙波 【张伟】 司成祥

    摘要：同一用户在不同社交平台注册账号，使得用户数据分散于多个平台，且这些数据不全面、不可靠、利用率低.通过分析这些跨平台的数据，发现不同账户对应同一用户的真实身份，使跨平台用户身份关联，以构建详细的用户画像、推荐系统、跨社交网络的链接预测等.从国内外身份关联技术的研究现状出发，介绍了用户身份关联及分析框架，整理了身份数据采集标准和社交网络数据集；分析了近几年用户身份关联技术，并归纳了身份关联评价指标，阐述了基于身份关联的社交网络数据挖掘及分析框架；最后对身份关联中的研究难点及热点进行了讨论和展望.

    关键词：跨平台 身份关联 身份识别

    URL：https://journal.bupt.edu.cn/CN/10.13190/j.jbupt.2019-020

    日期：2020-02-28


3. 相关度：50.0

    标题：多径中继传输系统网络仿真设计与实现

    作者：刘少伟 雷为民 【张伟】 付冲

    摘要：基于应用层中继的多径传输系统(MPTS-AR)是一种基于应用层中继的多径传输系统,通过基于用户数据报协议的中继服务器构建端到端的多径传输条件,并可支持多种业务实现多径传输. 提出基于OMNeT++平台上的INET Framework实现MPTS-AR网络仿真的方法,包括用户代理、中继服务器和中继控制器等网络逻辑实体的二次开发方法、相关网络链路和网络拓扑的构建方法、消息定义与处理方法等;讨论了应用程序接口兼容和多业务支撑等重难点技术. 最后,通过H.264业务传输实例验证了所设计MPTS-AR网络仿真的正确性和有效性.

    关键词：多径传输 应用层中继 应用程序接口 OMNeT++ INETFramework

    URL：https://journal.bupt.edu.cn/CN/10.13190/j.jbupt.2015.s1.021

    日期：2015-06-28


用户评价：
能很好找到作者
--------------------------------------------------
```

6.系统创新点与优化

6.1 创新性设计

1. 动态权重调节机制：
    - 使用Sigmoid函数将原始分数归一化到(0,1)区间
    - 通过立方IDF增强稀有词权重 IDF^3
2. 高频词过滤系统：
    - 双重过滤防止常见词干扰结果

    - 基于文档频率的预过滤层
3. 位置敏感评分：
    - 记录词语在文档中的具体位置
    - 不同字段赋予不同权重

6.2 可持续性优化

1. 资源效率：
    - 使用内存友好的字典结构存储索引
    - 分批处理大型数据集
2. 用户反馈机制：
    - 记录查询历史和改进建议
    - 高亮展示的搜索结果便于用户评估相关性