郑和后台计划

zcy

2019 年 10 月 10 日

第一部分

数据获取

第二部分

数据存储

第三部分

数据搜索

第 1 节　两大需求

Aminer 提供快速，精确的搜索服务。以学者搜索为例，系统会按照学者的论文数、被引用数、h-index 等指标排序并且能准确地区分**同名**的不同学者。

而在论文搜索中，Aminer 不仅仅能根据文章名进行搜索，还可以搜索与某一特定主题对应的论文。例如：直接搜索 *Data Mining* 就会出现数据挖掘领域的杰出论文。

因此，郑和平台目前最主要的两个需求如下：

- 根据关键词检索相关学者
- 根据关键词检索相关论文

考虑到用户体验和搜索结果的呈现质量，下述章节将基于这两大需求进行更为细致的分析，同时提供相应的解决方案。

第 2 节　关键词预处理

在查询数据库之前，至少应对用户输入进行如下处理：

- 对用户输入内容的容错（例如拼写错误或输入错误）
- 对同义词、近义词的处理
- 对词型变化及格位变化的处理（词干提取）
- 分词

上述大部分都或多或少地与 NLP 相关。在项目开始的初期（目前），我们考虑使用成熟的开源实现。

2.1　用户输入容错

从后端方向考虑，郑和平台所采用的图数据库内建对 Fuzzy Search 的支持，但需要建立相应的特殊索引。

2.2 同义词与近义词

基于 Word2Vec 的实现方式，进行同义词识别。

同时列出一些较为流行的开源实现以供参考：

- Synonyms[1]，预训练的中文相关词汇实现
- wordvectors[2]，支持多语言的 Word2Vec 实现

2.3 分词

英文语境下的分词依照空格分割之后去除介词和代词。

若来自用户的中文输入为句子（考虑到并非所有用户会用分隔符将关键词分开），首先应进行分词。

开源社区已有极为优秀和高效的中文分词实现 jieba[3]，具有如下优势：

- 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)
- 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合
- 对于未登录词,采用了基于汉字成词能力的 HMM 模型,使用了 Viterbi 算法

2.4 词干提取

相比分词，在中文语境下对用户输入进行词干提取的实用意义不是很大（直觉得出，尚无数据佐证）。

而在英文语境下，对关键词进行词干提取是必要的。例如，以关键词 “optimization” 和 “optimized” 分别进行搜索被期待返回接近或者相同的查询结果。事实上，Google 和 Bing 等搜索引擎确实对用户输入进行了词干提取。

采用 Porter 算法 [1] 进行词干提取效果较好，论文作者同时提供了免费的改进项目 Snowball[4]。

[1] https://github.com/huyingxi/Synonyms

[2] https://github.com/Kyubyong/wordvectors

[3] https://github.com/fxsjy/jieba

[4] https://snowballstem.org/

第 3 节　权重与优先级

对于每次搜索，用户输入经过预处理后应具有类似如下的结构：

$$i = \{p,\ p[\,],\ s[\,],\ r[\,]\}$$

p 为用户的原始输入字符串，$p[\,]$ 为分词结果，$s[\,]$ 为词干提取结果，$r[\,]$ 为相关词汇。

对 i 中每个元素赋予不同的权重 w_n，依据 $\sum w_n$ 决定搜索结果的相关程度。具体权重的分配方法仍需进一步讨论。

第 4 节　搜索词频统计

在数据库层面进行持久化，对每次搜索过程的分词结果 $p[\,]$ 和词干提取结果 $s[\,]$ 进行索引。

定期分析搜索日志，进行进一步分析（?）

第 5 节　热点数据缓存

为减轻服务器压力，应基于搜索词频统计结果，对热点数据进行缓存。

参考文献

[1] Porter, M. F. An algorithm for suffix stripping. *Program* (2006).