德塔 华瑞集 极速搜索算法基础

作者：罗瑶光

前言

- 很多朋友问我，为什么一个单机版本华瑞集软件能1秒处理200亿字搜索，我以前很想写一篇论文来描述这个算法机制，后来想想没有必要，ppt就足够了，本人感觉的小伙伴对ppt的理解能力绝对比论文要轻松。

总汇

- 1 主要成份 分析
- 2 关联成份 分析
- 3 重心成份 分析
- 4 极速分词 分析
- 5 带权评估 分析
- 6 有效过滤 分析
- 7 模糊数据 分析

主要成份 分析

- 1 德塔搜索的速度，确立在搜索的需求上，需求理解就是主要成份的确定，可理解为：
- 1.1 需求定义：定义需要搜索的话的表达方式。
- 1.2 需求预处理：将要搜索的话变成文字。
- 1.3 需求细化：这些文字进行文学组织精简练。
- 1.4 搜索的关键字：然后提取关键词语。

关联成份 分析

- 德塔华瑞集的关联成份分析，用于主要成份的相似度确认，和关联推荐。主要体现在
- 1 关联相似度：用于文章属性分类
- 2 关联聚类推荐：用于减少同类文章重分析
- 3 关联掩码：用于标识重分析的唯一性。
- 4 关联反码：用于对立类文章过滤和重分析
- 。

重心成份 分析

- 德塔养疗经的搜索，区分关键字和高频字以及重心字的定义:
- 1关键字: 作用于文章价值标识。
- 2高频字: 作用于文章重点标识。
- 3重心字: 作用于文章理解标识

极速分词 分析

- 分词的速度和准确性直接影响文本分析速度。
- 德塔目前的分词速度每秒可达1750万-2300中文词汇，中文分词精度99.7%。可支持每秒1300万的象形，契形混合70国，13个语种混合搜索，中文分词精度99.997%。
- 这个速度和精度，保证了德塔养料经华瑞集的极速分析能力。

带权评估 分析

华瑞集的搜索采用打分评估方式，权值在神经网络中的价值是巨大的。德塔的搜索采用5个维度进行耦合打分：

1 词性 通过词语的名动形介副 来加权

2 词长 通过词语的 单 双 三 成语来加权

3 词重 通过多个词语的邻接最小距离和来加权

4 词距 通过同一词语的欧基理德距离来加权

5 词感 通过词语的褒贬感情，词频来加权

有效过滤 分析

- 有效的过滤是筛选的重要主题。
- 华瑞集能在数百亿的文字中进行高准确率的筛选，过滤是非常重要的环节。主要体现在
- 1 明确过滤条件：用于判断过滤。
- 2 索引过滤条件：用于极速过滤。
- 3 微分搜索打分：用于精准过滤。
- 4 积分触发过滤：用于量变过滤。

模糊数据 分析

- 华瑞集在极速搜索出来的数据中，为了有效的辨识结果的准确性，进行了模糊统计算法处理。
- 1 隐词　　词汇 文章的隐藏属性。
- 2 比例词　　词汇 在文章中的出现比例。
- 3 同义词　　词汇的近义词和相似词。
- 4 概率差　　词汇的词性统计概率，词频统计概率 进行2次推荐。

华瑞集的搜索算法分为3部分介绍

- 1：用户---》搜索内容-》内容精简-》精简过滤-》关键字-》提出搜索请求。
- 2：服务器-》预处理-》索引文章-》格式化-》关联统一-》等待搜索。
- 3：搜索过程-》概搜（有效过滤，模糊数据，关联成份）-》细搜（主要成分，重心成分，带权评估）。
- 这个算法满足每秒200亿字 高精准确率搜索。

遇到问题

- 请加微信15116110525

- 谢谢！