**NBLucene论文研究报告**

1. A summary of the paper.

这篇论文主要讲述了一个扩展于Clucene的搜索引擎的实现，我们称之为NBLucene，其中扩展了几个部分，列举如下：

文本预处理（通过实现TREC web和Web Archive 功能的函数，主要功能为去除相关的HTML相关的标签，以及一些结束词，保留需要查询的主要内容），

索引压缩（与CLucene用的变长字节的压缩方法不同的地方为，NBLucene增加了以块的方式进行压缩，以及跨表随机查询posting list的部分），

执行查询（增加了Cache储存的机制，方便下次重复的查询，并实现了Cache的两种替换算法，分别为 LRU（最近最久未使用），LFU（最不经常使用））

返回结果（直接用了Lucene实现的BM25方法，进行相关文档的打分，然后返回K个相关结果）。

2.    A discussion of the main point/contribution of the paper.

本文最大的贡献即通过继承一个Clucene搜索引擎的架构，实现一些相关的接口以及函数，让人们意识到搜索引擎是可以根据自己的需要去灵活的扩展和修改。

3.   A discussion of the strengths of the paper.

1： 首先是语言，由于Java语言的几个缺点 :

I：运行Java虚拟机上，速度较慢；

II：垃圾回收和内存的分配不受程序员控制

III：Java语言很难这种并行的指令SIMD 和 支持并行的GPUs的平台.

而通过C++实现的搜索引擎能够很好克服这两个缺点。

2： 灵活的集成了自己需要搜索引擎需要的东西，让搜索引擎更加完备。

I:文本预处理，由于Clucene并没有直接提供任何关于HTML标签处理的工具，自己通过实现相关的函数，比如说去掉HTML和Js的相关标记，把相关的文档主要元素（比如说URL，Title，Content）进行重新编排，转换成标准的doc文档进行下一步处理。并且可以让用户选择看是否过滤掉一些无意义的词汇(即停止词)

II:在NBLucene文中，压缩Posting list可有效的节省资源。

由于每个文档在搜索引擎中被重新编为唯一的docID,并且在实际倒排列表中，存储docID时进行的是 差值存储，（因为docID可能会很大，而相关联的差值可能会很小）。在这里，主要将docID和倒排索引中相关的词频进行压缩。freq（词频）如果为1，利用docID的左移(末位补1)即可。

如果不为1，docID后面需要跟实际的freq。

以1bit 换 1个字节 实际效果在大量数据前还是十分可观的。

这里无论freq为1 或者 > 1都进行左移，主要是为了需要留出一位，来进行标志，看是否需要后面是否有这个必要跟随这个实际的词频。主要的目的就是将来解压的时候，将有压缩的和没有压缩进行区分。将来解码的时候可以用if else进行判断。

III:在解码的时候，通过一些并行指令（SIMD），可实现压缩数据的迅速解压。（通过复制和相关的移位，然后进行与运算，即可迅速解压出原来的内容）。

IV:利用缓存Cache进行存储相关的查询结果，并且实现Cache的两个相关替换算法，（LRU最近最久未使用和LFU最不经常使用），提高了重复查询的效率问题。

4.   A discussion of the weakness of the paper. If none, why not?

1: 从结构上看，并没有发现什么问题。。理由如下:

首先通过先简介现在存在的搜索引擎存在的问题，比如说为了跨平台的使用，运用Java语言进行编写，然后指出Java语言的两个弊端，提出想要改写搜索引擎。

然后介绍相关工作，通过扩展四个方面，即（文本预处理，索引压缩，执行查询，和返回结果）并且进行一一介绍。

其次对倒置索引压缩和并行计算解压缩进行详细介绍。

接下来对增加的Cache部分进行了详细介绍，并写出了LRU和LFU的基本原理。

最后就是根据相关内容进行的不同实验以及相关总结。

我觉得从结构上来看，基本没有什么缺失的部分，对于比较复杂的倒置索引压缩和解压缩相关内容也进行了较为详细的介绍。

2：从内容上，由于我是初学者。。刚接触搜索引擎这一块，。我只能凭我自己的理解去找内容上的相关缺点。。。如果有不对的地方，还望老师多多指教。

I:由于我是搜了一篇中文的硕士论文，来进行入门搜索引擎的，那篇中文论文详细介绍了中文检索的相关内容，对按词还是按字进行划分来建立倒排索引进行一定程度的讨论。而这篇NBLucene 并没有看到相关的内容的讨论，所以我认为，在内容上说这可能是一定程度上的缺失。

II：这篇论文可能不太适合入门搜索引擎的同学。有省略较多的地方，就比如说索引建立这一块，直接说After Index building，相关内容基本就跳过了，直接讲解索引压缩的相关知识。我觉得如果可以的话，可以稍微提一下索引的基本概念，然后给出几篇比较好的介绍索引的基本知识的相关论文，这样，即使是初次接触搜索引擎，也可能知道上哪去找相关的名词概念。还有一个体现，最后说未来的希望的相关工作时，直接就提出了WAND查询，连个参考论文资料都没有，这让初次接触搜索引擎的人，可能会比较茫然。

这内容上的相关缺点，可能是我作为初次接触搜索引擎的人都会遇到的，我也没有相关读论文的经验，如果有说的不对的地方，还请学长和老师们多多指教。

5.   A discussion of what you learned from the paper.

首先非常感谢老师给我推荐的这篇论文，在阅读的过程中，基本上没什么障碍，主要大致意思都能够弄懂，除了个别专业词汇和相关的方法在初次接触中有点困难，但经过老师和学长的细心指点，基本已经能够顺利阅读。在此也再次感谢老师和学长的耐心指导！ 我也从两个方面来谈我从这篇论文中学到了什么。

知识上说：大致知道了倒置索引，索引列表，索引压缩，解压列表等搜索引擎基础知识有了一定的了解。

入门了搜索引擎，大致了解了用户输入想要查询的词汇后，后台搜索引擎将要以哪些步骤就行处理。我现在考虑到的步骤大致为：用户输入搜索关键词 -> 文本进行预处理，提取关键词并保留有逻辑AND 或者 OR 或者 NOT等等 -> 去已建立好的索引进行查询，（建立索引时需要压缩的知识，了解到了几个压缩的算法，比如说 变字长压缩（VByte），NewPForD，Simple9压缩算法等，还有相关的解压缩的知识，比如说通过判断移位后的docID末位标志位是否为1来判断是否进行了压缩等等）-> 搜索引擎进行相关内容的排序 –> 返回给用户查询的根据优先级进行排序的相关文档。

另外对SIMD并发解压缩的步骤和Cache缓存的替换策略有了一定的了解。

非知识上的：第一次这么认真的读一篇文章，遇到困难的通过查资料，与老师交流，然后不断思考，最后得出肯定的答案。我觉得这个不断去找问题的答案的过程，或者说自己思考后得出答案的过程，是我看这篇论文最大的乐趣，最后也成为我最大的收获。

在这里再次感谢王刚老师对我的耐心指导和曌华学长对我疑惑的解答 !