信息检索课程总结

71Y17102黄少豪

* **课程心得**

在信息检索这门课程中，我学习了搜索引擎的底层原理，明白了面对海量数据的信息检索，不同于数据库简单的查询，而是一个从信息集合中找出所需要信息的匹配过程。而检索效率的关键，在于检索模型的构建，以及信息存储的方式。

* **倒排索引**

区别于常见的正向索引（forward index），倒排索引（inverted index）不是通过文档去找关键词，而是通过关键词去匹配包含它的文档，这符合我们的搜索逻辑，若采用正向索引，则每次查询都需要遍历所有文档，效率太低。因此，它是搜索引擎中非常重要的数据结构。

* **布尔模型**

布尔模型查找返回为真/假，一个查询词就是一个布尔表达式，这种模型非常直观，查询语义可以非常精确，但缺点也很明显：缺乏文档分级概念，只有相关和不相关两种文档，会导致过多或过少的文档被返回；用户查询需求往往不是很容易能转换成布尔查询式，这些限制了布尔模型的应用场景，往往是文档数据库系统。

* **向量模型**

为了改进布尔模型非真即假的局限性，向量模型提出了适合部分匹配的权重和排序概念。查询语句不再有AND, OR的概念，而是以向量的方式存储。词袋模型将文档转化成一个个词，构成整个词袋，这导致文档向量往往很稀疏。同时引入tf-idf概念，计算各个term项的权重。最后根据余弦相似性算法，计算每篇文档与查询语句的匹配程度，即得分，并进行排序。向量模型根据秩（rank）返回的结果集通常比布尔模型返回的结果集在感觉上更加符合检索用户的需要。

* **概率模型**

概率论模型的基本准则是：最优的文档排序应为按照文档与当前查询相关的概率降序排序。PRP（probability ranking principle）：概率排序原理，利用概率模型来估计每篇文档和需求相关概率，然后对结果进行排序。贝叶斯最优决策原理，基于最小损失风险作出决策，返回相关的可能性大于不相关的可能性的文档。二值独立模型BIR，目标在于估计查询qk，文档dm被判定为相关的概率，最基本的假设为词项在相关文档与不相关文档中的分布是不同的，如果相同，则他们的得分将会完全一样。

* **有效存储**

为了最大化数据传输速率，一起读取的数据块应在磁盘上连续存储。对于小的倒排索引集合，可在内存中完成。而对于大型集合，内存无法满足时，需要改进存储效率。

* 基于块的排序索引（Blocked sort-based indexing）

将文档解析为termID-docID对，并将它们在内存中积累，直到固定大小的块被填满。选择块大小来适应内存，以允许快速的内存排序。步骤包括切割集合、键值对排序、存储和合并结果。

* 分布式索引（Distributed indexing）

当数据集庞大到无法在一台机器上高效存储索引时，我们就要使用分布式索引算法进行索引构建。比如按照A~Z的字典顺序分布式存储，查询时只需去指定的区域就能快速匹配，如同图书馆中对书的有序排放。实现方法可以用MapReduce分布式架构。

* 动态索引( Dynamic indexing)

针对需要频繁修改的，有文档被添加、删除和更新的集合，动态索引通过同时维护两个索引，即大型主索引和存储新文档的小型辅助索引，以及一个已删除文档列表，用来存储已被删除的文档的相应文档ID，满足实时性的要求。实现过程：当系统发现有新文档进入时，立即将其加入临时索引中。有文档被删除时，则将其加入删除文档队列。文档被更改时，则将旧文档加入删除队列，解析更新后的文档内容，并将其加入临时索引中。

* **实验心得**

在每一章节结束后的实验同样让我受益匪浅，实践书本上的模型原理不仅锻炼了代码能力，更加深了我对模型原理的理解，每次实验结果也让我对各种模型优劣及其适用的情况有了更深的理解。

从一开始爬取歌曲信息、构建倒排索引，到实现最简单的布尔模型，再一步步慢慢优化，优化查询界面、加入k-gram、zone index提高查询效率、实现向量模型、概率模型，最后看到自己的查询语句与目标匹配，是一件很有成就感的事。

其中，我感触最深的是向量模型。在实现了向量模型的三种不同得分算法，两种余弦相似性算法，进行对比阶段，我发现得分前三的有时并不是预期目标。影响结果的因素有很多，比如有些停用词没加，导致无用的权重占比过高，有些是文章长度较短，导致得分偏高。在改进了这些问题后，除未归一化的tfidf得分方式匹配程度仍然欠佳外，其他两种算法都能较好的匹配目标结果。余弦相似性算法FasterCosine和Cosine方面，对比结果显示区别不大。实验中还出现了一些小bug，比如没有处理输入为空、查询为空的情况，这提醒了我程序健壮性也很重要。

* **搜索引擎的未来**

我认为未来的搜素引擎会与AI技术融合，在AI的加持下为我们提供更智能化的搜索功能。其中我对知识图谱比较感兴趣。

目前知识图谱的应用主要在搜索引擎，提高搜索质量，给出结构化结果，省去了一次次查询。虽然现在知识图谱的应用场景还较少，甚至查询结果还需要人工干预以保证正确性，但其未来重要性不言而喻，未来智能终端想要拥有“大脑”，知识图谱提供的知识库不可或缺。但就目前而言，仍然有许多需要解决的问题：知识表示方面，三元组虽然能较好的表示事实型知识，但对人类更复杂的知识、模糊的主观感受、更复杂的实体则有些无力；知识获取、知识融合方面，如何提高准确率，萃取有效信息，减少冗余和噪声，是构建更大规模知识图谱的基础；如何提高推理学习的能力，使得机器能够“举一反三”等等。

我相信未来的知识图谱，对于大数据智能具有重要意义，对自然语言处理、信息检索和人工智能等领域将产生深远影响。