从算法角度分析两类情报检索系统*

张 进 陈 远

作 者 张进，武汉大学图书情报学院教授，武汉，430072;
陈远，武汉大学图书情报学院讲师，武汉，430072

关键词 脱机批处理系统算法 联机检索算法 比较研究

提 要 文章从算法角度对联机情报检索系统与脱机批处理系统进行了讨论，分析了它们在对逻辑非算子的处理、检索周期、处理最佳终止点、比较时主次方、文档结构，以及提问逻辑式机内等价形式对效率影响等方面的差异。

脱机检索系统与联机检索系统是情报检索系统的两大主要类型。在了解两大类型检索系统的算法运作及原理的同时，从算法处理角度对它们进行对比性分析研究，这对从理论上全面掌握其工作原理以及于实践中将其进一步改进，都是极有价值的。

虽然脱机检索系统与联机检索系统都是对用户提问式进行处理，并给用户提供命中文献结果，但由于二者目的不同，处理对象的数据结构及内部处理机制也不同。为了加深对两大检索系统算法的本质理解，我们抽出其核心部分，进行对比分析。

1. 两种算法对提问式的处理不同。我们知道，任意一个提问逻辑式都可以经过一定的变换转化为标准的析取范式: $A = B_1 + B_2 + \cdots + B_i + \cdots + B_n$ (B_i, $i \leqslant n$是一子合取式)。

脱机批处理检索算法，无论是菊池敏典算法，还是欧美算法，或是改进 SDI算法，在算法的执行过程中均是以提问词与文献记录中标引词进行匹配。只要提问表达式中有一个子合取式成立，即该子合取式中各检索词与一个文献记录中的标引词分别匹配成功，那么，算法就可以中止。这是因为在算法的一个处理周期中，它只是鉴别一个文献记录是否满足该提问式，而对联机检索算法来讲，它的每一个处理周期中需要确定的则是所有数据库中满足提问式的一批文献记录。因此，它必须对提问逻辑式中所有可能满足条件的子合取式进行处理，以确定一组文献记录。

2. 算法执行过程中最佳终止点的确定问题。脱机批处理菊池敏典算法展开表技术的一个最大特点，就是在算法执行中对提问式中各提问词的匹配处理一旦满足终止条件，决不多做

* 本论文受国家社会科学基金 (青年) 项目资助

任何多余的匹配工作，立即终止算法的执行，输出结果。正如前面所指出的那样，一旦有一个子合取式成立，再对其它子合取式中检索提问词的继续匹配处理就是多余的。因此，在脱机批处理算法中存在着匹配过程最佳终止点的选定问题，这是对脱机批处理算法的一个特殊要求，联机检索算法不存在这种问题。

3. 算法的扫描执行次数问题。在脱机批处理系统中，如果被检索的顺序文献数据库中有 N 篇文献记录，那么，算法将执行 N 次。这是因为脱机批处理算法在一个周期的执行过程中，提问文档每次仅为一个文献记录中的标引词进行匹配处理，以决定该文献记录的取舍。联机检索算法在对提问式处理时，仅执行一次，每次不是针对一篇文献记录。这是因为倒排文档中，每个检索提问词后所附属的文献地址集合，是针对整个文献数据库的，一次性处理后，就可以决定最终命中文献集合。

4. 检索时主次比较方的确定问题。在检索时，提问文档中的提问词与文献主文档中标引词进行比较，这里存在着比较双方的主次问题。在脱机批处理算法中，既可以以提问文档为比较的主方，文献主文档为比较的次方；也可以以文献主文档为主方，以提问文档为次方。由于提问文档和文献主文档在比较过程中所处的主次地位不同，由此可以划分为欧美算法流派和东方算法流派（这主要是指菊池敏典算法）。在联机检索执行过程中，只能以提问文档一方为比较时的主方，倒排文档一方为次方。联机检索在比较时文献主文档一方就是整个倒排文档，其入口词的数目是相当大的一个数 N。假设以倒排文档为比较时的主方，提问文档一方为次方，并且同时假设提问文档中有 m 个提问词，显然有 $m \ll N$ 次。若按顺序搜查找方法，对整个提问式的处理要比较 $m \times N$ 次。以提问文档为主，由于整个倒排文档的入口词已进行排序处理，在搜集时可以用二分查找算法，其比较次数为 $m \times \log_2 N$。两者比较起来，显然后者比较次数要少得多。这就是为什么联机检索算法中以提问文档为主方的原因。脱机批处理算法处理中，算法每执行一个处理周期，提问文档仅与一个文献记录中的标引词进行比较，它们的数量是有限的，比起整个文献数据库的文献量要小得多。因此，以提问文档或文献主方档中任何一方作为比较时的主方都不妨大局。

5. 提问式中提问词倒排文档的建立问题。在脱机批处理算法的执行过程中，提问式往往是成批的。在各提问式中很可能存在相同的提问词。无论是以提问文档还是以文献主文档为比较时的主方，建立这样的提问词倒排文档，可以减少提问词与文献标引词的匹配次数，提高算法效率。联机检索软件一般只处理一个提问式，并且对各提问词是顺序执行的，没有建立提问词倒排文档的必要。

6. 为了提高查找效率，联机检索的主文档—— 倒排文档，由于词最大，往往配有字顺索引，这样做虽然用去了一些空间，但可大大提高查找速度。脱机检索系统没有建立这种索引的必要。这是因为如果要建立这样的索引，必须给每篇文献记录中的标引词建立索引，建立这样的索引对提高速度作用不十分大，相比之下，又占用大量空间，显然是不合算的，故一般情况下脱机检索系统的主文档各记录内数据不建索引。

7. 文献文档结构的差异。脱机检索系统中的文献组织结构是顺排结构。它以每一篇文献为组织信息的单元。在一个文献记录之下，按一定方式将描述该文献内容特征、外表特征的所有标引项集中加以管理。而在联机检索系统中，每篇文献下各标引词分散在数据库中各处加以管理。在这样的结构下，按每一个标引词为组织信息的基本单元，在每一个有检索意义

的标引词后，附属着所有相关文献号，即文献是按某一特征加以集中的。

8. 两种检索系统对提问逻辑式格式限制不同。脱机检索系统和联机检索系统均采用布尔逻辑作为提问表达的手段，然而两者在使用布尔逻辑作为信息提问表达手段时，均有一定的局限，不能完全等同于布尔逻辑。一般情况下，脱机检索系统在使用逻辑算子时，逻辑非只能作用在提问词上，不能作用在子表达式上。

联机检索系统对逻辑非的使用也有一定的限制，它不允许逻辑非以及它所作用的因子作为析取因子出现在提问式中。例如: A+ NOT B+ C是不允许的; 有的系统不允许逻辑非以及它所作用的因子出现在提问式的开头。例如: NOT A* B+ C, 上式必须转化为 B* NOT A* C或 B* C* NOT A后方能处理。

9. 两种检索算法处理对象的范围不同。从宏观上看，脱机批处理算法在整个算法执行完毕以后，对顺排文档中所有的记录均进行了一次查询，其检索结果是针对整个数据库文献记录的。但是具体到脱机检索算法的检索过程，每一个处理周期只能确定一个特定记录是否与一个特定提问之间满足检索条件，然后决定是否将该记录放入命中文献文档中。这样的处理过程从顺排文档的第一记录开始一直进行到顺排文档的最后一个记录为止。

与脱机检索系统在此问题上的一个明显差异是，联机检索算法在每一个处理周期确定一个命中文献记录的集合，而不是一个文献记录。进行操作的集合个数取决于提问表达式的提问词个数。由于倒排文档中每一个入口款项，也就是每一个关键词所对应的文献记录号集合是相对于整个文献记录的，因此，最终的集合运算结果也是针对整个文献记录而言的。从这个意义上讲，联机检索在检索时的高效率是以建立倒排文档的额外存贮开销为代价的。

10. 脱机批处理算法和联机处理算法在将提问逻辑表达式转化为机器等价形式的过程中，对检索时间效率的影响是截然不同的。脱机批处理算法，无论是菊池敏典算法，还是欧美算法，从整个检索过程上看，可以分为两大步骤: 第一步，算法将用户提出的各种提问转化为机器内部的等价逻辑形式，例如菊池敏典算法的提问展开表等。这些机器内部的提问逻辑式的等价形式一旦形成以后，可以长期保存，多次使用。也就是说在实际检索开始之前，它们已经形成并存贮在计算机内部，第二步检索时，再调用它们。因此，脱机批处理算法中将提问表达式转化为机器内部等价形式的过程不直接影响检索时间效率。

对于联机检索算法来讲，情况则大不一样。由于联机检索在检索过程中，提问式形式的转化以及检索过程都是实时进行的，因此，第一步将提问式转化为机器内部等价逻辑形式的过程和第二步利用它们进行检索的过程是合并在一起的,两者的时间之和便是整个检索时间，也即将提问逻辑表达式转化为机器内部等价形式的过程将直接影响联机检索的时间效率。

参考文献:

1 张进:《对菊池敏典算法逻辑非处理的讨论》，载《现代图书情报技术》1991年第 2期。

2 张进:《布尔逻辑与提问逻辑的差异分析》，载《图书情报知识》1992年第 4期。

(责任编辑 张 琳)