ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.1, January 2008, pp.136−143
DOI: 10.3724/SP.J.1001.2008.00136

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

分布式信息检索中文档集合划分问题的评价*

张 刚[1,2+], 谭建龙[1]

[1](中国科学院 计算技术研究所,北京 100080)

[2](中国科学院 研究生院,北京 100049)

Document Collection Partition Evaluation in Distributed Information Retrieval

ZHANG Gang[1,2+], TAN Jian-Long[1]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-62600988, Fax: +86-10-62600905, E-mail: gangzhang@ict.ac.cn

Zhang G, Tan JL. Document collection partition evaluation in distributed information retrieval. *Journal of Software*, 2008,19(1):136−143. http://www.jos.org.cn/1000-9825/19/136.htm

Abstract: It is difficult to evaluate the document collection partition in distributed information retrieval. Recently, there is no clear evaluation criterion for the document collection partition problem. In this paper, two partition models are built to formulate the document collection partition problem from the essence of the problem itself and they can be used as the evaluation criterion of the document collection partition problem. A Huffman_encoding_like algorithm is introduced to compute the optimum partition solution given a test query set. The optimum partition solution is a good reference of other partition solution. The experimental results show that the two models are effective document collection partition evaluation criteria.

Key words: distributed information retrieval; document collection partition; Huffman code

摘 要: 分布式信息检索的文档集合划分方案的评价是一个困难的问题,目前还没有良好的评价标准.从文档集合划分问题本身出发,给出了两个划分模型来刻画文档集合划分问题,从而使这两个模型可以作为文档集合划分的有效评价指标.在此基础上,提出了一种类 Huffman 编码的模型快速求解算法,可以求出在给定查询测试集情况下的最优文档划分方案,该方案可以作为其他文档划分方案的参考.实验表明,两个文档划分模型可以成为有效的文档集合划分评价标准.

关键词: 分布式信息检索;文档集合划分;Huffman 编码

中图法分类号: TP311 **文献标识码**: A

近年来,Web 信息增长迅速,尤其是当博客、网络论坛、个人主页等动态网页出现并流行时,更使 Web 信息达到一个空前的规模.据 Hobbes' Internet Timeline 统计,截止到 2005 年 8 月,互联网上 Web 服务主机数已达到 70 392 567 台,著名的搜索引擎 Google 声称索引的页面数已超过 80 亿,而这也只是全部网页中很少的一部分.迅速增长的 Web 信息给检索系统带来了巨大的挑战.传统的集中式搜索引擎对于每个查询都搜索全部文档集

* Supported by the National Basic Research Program of China under Grant No.2004CB318109 (国家重点基础研究发展计划(973))
Received 2006-07-08; Accepted 2006-11-21

合,当数据规模很大时,其计算开销也相应增加.而分布式信息检索能够通过仅搜索部分文档集合,就可以给出检索结果,是海量信息检索的有效解决方案[1].分布式信息检索主要包含“集合划分”[2]、“集合选择”[3,4]、“单数据集合检索”、“结果合并”[5,6]等几个部分.

“集合划分”是构建分布式系统的第 1 步,也是至关重要的一步,因为文档集合划分的好坏直接影响到分布式信息检索的质量[1,2].在以往的研究中,对于文档集合的划分常常采用一些启发式的信息[2],例如,将相同主题的文档划分在同一个集合里.但对于文档集合划分的原则没有明确的论断,尤其是在文档集合划分的评价方面没有统一的标准,使得不同的划分方案不能合理地进行比较.由于没有专门的评价方法,在以往的工作中,对于文档集合划分的评价通常采用最终分布式信息检索结果的质量来评价文档集合划分的好坏,但这种评价是一种间接的评价,最终分布式信息检索结果的优劣不仅取决于“集合划分”,还受到“集合选择”、“单数据集合检索”以及“结果合并”的影响.因此,这种方法对于文档集合划分的评价是不精确的.

本文从文档集合划分的基本问题出发,针对文档集合划分问题进行建模,从理论上给出了一个最优划分方案.通过这两个模型,可以对文档划分问题进行有效的评价,同时,按照该模型可以计算出给定查询测试集的最优的划分方案.

1 文档集合划分的问题描述

分布式信息检索的过程可以如下描述:将文档划分成若干子集合,在检索时,首先找到最相关的部分子集合(集合选择),再对这部分子集合进行检索,找出其中的相关文档(单数据集合检索),呈现给查询者.分布式信息检索的集合划分问题,实际上要回答这样两个问题:一是文档应该被划分为多少个子集合;二是每个子集合应该包含哪些文档.直观的想法是,一个查询的相关文档应该分布在一个或尽量少的数据集合中,这样,检索时才能尽量减少不必要的查找.

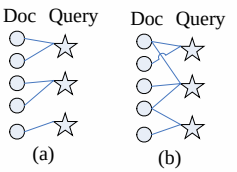

Fig.1 The relationship between document and query

图 1 查询与文档的对应关系

如果文档和查询之间是“一对一”或者“多对一”的关系,如图 1(a)所示(连线表示相关性),那么划分问题非常简单,只要将同一个查询的相关文档放入一个子集合中就可以了;但文档和查询之间往往是“多对多”的关系,如图 1(b)所示,此时,要保证同一个查询的相关文档都处在同一个文档集合中就可能会出现冲突的情况,此时,就要寻找一种更好的划分方法.

2 文档集合的划分模型

2.1 文档集合划分的优化目标

通过对文档集合的划分,我们希望分布式信息检索系统能够用最少的计算开销获得最好的检索结果.这是一个多目标规划问题,既要检索的结果好,又要检索的文档数量少.要使检索的结果好,那么,被选择来进行检索的文档集合中一定包含全部查询的相关文档;另一方面,要使检索的文档数量少,就需要被选择来进行检索的文档集合含有尽可能少的不相关文档.为了更好地表达问题,我们定义“待查询文档数”和“平均待查文档数”这两个概念.

定义 1(待查询文档数). 待查询文档数是指在给定一种文档集合划分方案的条件下,对于一个查询,包含其相关文档的各文档集合所含的文档数之和.

定义 2(平均待查询文档数). 在给定一种文档集合划分方案和一组查询的条件下,对于每个查询,计算其“待查询文档数”,平均待查询文档数是指所有查询的待查询文档数的算术平均值.

显然,对于每个查询,我们都希望其“待查询文档数”能够达到最小,“待查询文档数”的最小值就是该查询的相关文档数.但是,通过上面的文档集合划分问题的描述可知,在实际操作中,使每个查询的“待查询文档数”达到最小是不可能的.因此,文档集合划分的优化目标就是通过变换文档集合的划分方案,希望给出一种“平均待查

询文档数”最小的文档集合划分方案.

2.2 文档集合划分优化问题建模

2.2.1 文档集合划分模型 1

在分布式信息检索过程中有两个阶段的检索操作:第 1 阶段是集合选择过程,这一阶段文档集合被看作是一个虚拟的文档,采用与文档检索相同的算法,检索到相关的文档集合;第 2 阶段是对被选择的文档集合进行检索,找出与查询相关的文档.我们建立下面的模型来刻画数据集合划分问题,令 $Q=\{q_1,q_2,\ldots,q_n\}$为用户输入的查询集合,R_j 为查询 q_j 的相关文档集,K 为文档集被划分的子集合的个数,$S_i(i=1,2,\ldots,K)$为第 i 个子集合.在检索时,首先要在 K 个子数据集中选择出含有相关文档的子集合,这些子集合满足:$S_i\cap R_j\neq\varnothing$,然后再对这些子集合分别进行检索.因此,对于查询集合 $Q=\{q_1,q_2,\ldots,q_n\}$中的每个查询 q_i,要检索到其全部的相关文档,需要查询的平均文档总数为(这里文档集合也被看作虚拟的文档):

$$avgdoc1=\frac{\sum_{q_j\in Q}\left(K+\sum_{S_i\cap R_j\neq\varnothing}|S_i|\right)}{|Q|} \tag{1}$$

上式为划分的子集合个数不确定情况下的文档集划分模型,称为划分模型 1.模型 1 是希望目标函数值 $avgdoc1$ 的值最小,此时,模型参数 K 和一组子集合 $S_i(i=1,2,\ldots,K)$就是集合划分问题在给定查询集 Q 下的最优方案.

上面的模型是基于集合选择和文档检索的检索算法为线性的条件做出的.事实上,如果文档的索引采用倒排表结构,搜索时是布尔模型,容易验证算法的复杂度是线性的,“集合选择”时会把子集合当作虚拟文档建立索引并检索.因此,“集合选择”的复杂度也是线性的.

进一步考虑模型 1 在两种极端情况下的表现.在传统的集中式检索中,没有数据集合的划分,此时,模型 1 表示为

$$avgdoc1=\frac{\sum_{q_j\in Q}(0+N)}{|Q|}=N \tag{2}$$

其中,N 为文档的总数,每次查询时,没有“集合选择”的阶段,直接面对全部的文档进行检索.

另一种极端的情况是,每一个文档都作为一个文档集合,此时,模型 1 表示为

$$avgdoc1=\frac{\sum_{q_j\in Q}(N+0)}{|Q|}=N \tag{3}$$

在这种情况下,“集合选择”直接查询出来的子集合就是相关文档,不必再对每个子文档集合进行检索.

这两种情况在实践中是一样的,而按照上面的模型,它们的 $avgdoc1$ 值也是完全相同的.

2.2.2 文档集合划分模型 2

在划分的子集合数量 K 确定的情况下,划分主要和文档在各个子集合中的分布有关.因此,这时模型可以简化为

$$avgdoc2=\frac{\sum_{q_j\in Q}\sum_{S_i\cap R_j\neq\phi}|S_i|}{|Q|} \tag{4}$$

式(2)是子集合个数确定情况下的划分模型,称为划分模型 2.模型 2 所表示的就是“平均待查询文档数”.

2.2.3 文档集合划分模型 3

上面的模型是在查询出现的概率是均匀分布情况下得出的,而实际查询出现的概率并不一致.如果考虑到查询 q_j 出现的概率为 p_j,则模型 1 可以进一步表示为

$$Score=\frac{\sum_{q_q\in Q}p_j\left(K+\sum_{S_i\cap R_j}|S_i|\right)}{|Q|} \tag{5}$$

在考虑到查询出现概率的条件下得到的模型称为模型 3.

3 模型的可行解分析

将小球放在盒子里的过程可以用来形象地说明文档集合的划分问题.小球可以比作文档,而盒子则代表文档集合,对于划分的子集合数量不确定的情况,可以描述为要把 N 个有区别的小球放在 N 个没有区别的盒子中,允许有空的盒子.在所有的可能中找出一种最好的放置方法,使得模型 1 的取值最小.由组合数学知识可知,所有可能的放置方法总数是

$$S(N,1)+S(N,2)+\ldots+S(N,N),$$

其中,$S(N,i)$为第 2 类 Stirling 数.

对于划分的子集合个数确定的情况,可以描述为把 N 个有区别的小球放置在 K 个没有区别的盒子中,不允许有空的盒子.那么,在所有可能的放置方法中寻找一种最优的放置方法,使得模型 2 的取值最小.所有可能的放置情况为 $S(N,K)$.

4 文档集合划分问题最优解的快速解法

由上面的模型定义可知,模型 1 与模型 2 之间有如下关系:$avgdoc1=K+avgdoc2$,如果求出 $avgdoc2$ 的最优解,则只要再将 K 扫描一遍,就可以求出 $avgdoc1$ 的最优解.由解空间分析可知,模型 2 的可行解空间是第 2 类 Stirling 数,当文档数很大且划分的子集合个数很多时,可行解空间非常大.因此,采用穷举的方法理论上是可以求得最优解的,但实际中却是不现实的.

下面给出一种针对于两个模型的一种快速求解算法,称为类 Huffman 编码的文档划分算法(Huffman_encoding_like partition algorithm).

4.1 模型2与Huffman编码问题的类比

在这种快速解法中,模型 2 被等效为一个 Huffman 编码问题,下面比较一下 Huffman 编码问题和模型 2 的文档划分问题.在 Huffman 编码中,有 3 个重要的概念:字符、字符编码长度、字符出现的频率.字符就是等待被编码的字符;编码长度表示用多长的二进制编码表示这个字符;而字符出现的频率是 Huffman 编码中影响字符编码长度的因素.按照字符出现的频度给字符不同长度的编码,出现频度大的字符采用短的编码,出现频度小的字符采用长的编码,这样可使报文中码数减少至最小.

对于模型 2,文档集 D 被划分为 K 个子集 $S_i(i=1,2,\ldots,K)$,令 S_i 就是一个要编码的字符,而子集合的模$|S_i|$为该子集合的字符编码长度.如果对于查询集 $Q=\{q_1,q_2,\ldots,q_n\}$检索 Q 中的每一个查询时,文档集合 S_i 被使用的次数就是字符 S_i 出现的频率,表示为$|T|$,其中,T 表示为 $T=\{t_j|S_i\cap R_j,j=1,2,\ldots,n,$且 $t_j\neq\varnothing\}$,$R=\{R_1,R_2,\ldots,R_n\}$为查询集对应的相关文档集.

经过这样的类比变换,计算模型 2 的 $avgdoc2$ 最优值就是希望找出一种与之对应的 Huffman 编码.最优解的构造过程采用自底而上构造一棵 Huffman 树,初始状态下,每个文档都是一个独立的文档集合,进而选择两个文档集合合并,合并的过程与 Huffman 编码相同.但在模型 2 计算字符 S_i 出现的频率$|T|$时,如果$|T|=1$,我们规定此时字符出现的频率为 0,因为此时该子集合的所有文档都是同一个查询的相关文档,没有出现一个文档属于多个查询的情况,不会影响到 $avgdoc2$ 的结果.

4.2 类Huffman编码快速求解算法

具体的划分过程可以模拟 Huffman 树的构造过程,算法可以描述如下:

令 j 表示每次迭代中划分的子集合个数,k 表示第 k 次迭代,N 为全部文档总数.

(1) 初始条件下每个文档为一个子数据集合,此时,$j=N,k=1$.

(2) 分别计算两个模型的最优值:

$$avgdoc2^k = \sum_{q_j \in Q} \sum_{S_i^j \cap R_j \neq \varnothing} |S_i^k| \Big/ |Q|,$$

$$avgdoc1^k = j + avgdoc2^k.$$

(3) 计算第 k 次迭代中任意两个子集合 S_x^k, S_y^k 的合并代价,计算方法如下:

令

$$S' = S_x^k \cup S_y^k,$$

$$T=\{t_j | t_j = S' \cap R_j, j=1,2,\ldots,n, \text{且 } t_j \neq \varnothing\},$$

$$cost=|S'| \times |T|.$$

取合并代价 cost 值最小的两个集合 S_x^k, S_y^k 进行合并(如果有多个最小值相同的情况,可以任选一组进行合并),将合并后的集合 S'取代原来的两个子集合 S_x^k, S_y^k,形成一种新的集合划分方案,此时,子集合个数少了一个,即 $j=j-1$,迭代次数 $k=k+1$.

(4) 如果子集合个数 $j>1$,则跳到(2)步继续执行

(5) 通过上面的计算,可以得到一组 $avgdoc2^k$ 和 $avgdoc1^k$,其中,$avgdoc2^k$ 为第 k 次迭代时模型 2 的最小值,其所对应的一组子集合 S_i^k 为模型 2 的最优解.

(6) 扫描一遍 $avgdoc1^k$,通过比较求出 $avgdoc1^k$ 的最小值 $\min(avgdoc1^k)$,它所对应的一组子集合 S_i^k 为模型 1 的最优解.整个构造过程如图 2 所示.

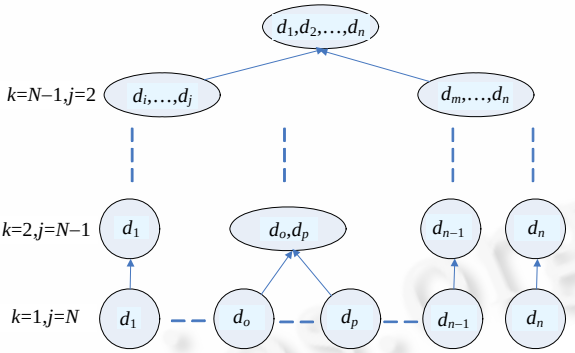

Fig.2 Construction of the optimum solution

图 2 模型最优解的构造示意图

4.3 对于最优解的几点讨论

(1) 在对文档集合 D 进行划分时,如果 D 中的某些文档没有出现在任何一个查询的相关文档集合中,那么这些文档构成子集合 D',$D'=\{d_i | d_i \in D \text{ 且 } d_i \notin R_j, j=1,2,\ldots,n\}$.因为 D'与任何查询都不相关,因此,可以令 D'作为一个子集合,或者直接从 D 集合中去除掉,而不会对模型产生影响.

(2) 如果在查询的相关文档集合 $R=\{R_1,R_2,\ldots,R_n\}$中,对于任何 R_i,R_j,都有 $R_i \cap R_j=\varnothing$且集合 R 的模$|R|$小于要划分的子集合个数 K;或者相关文档集合 R 可以被拆分为 $R'=\{R'_1,R'_2,\ldots,R'_m\}$ 满足对于任何 R'_i, R'_j,都有 $R'_i \cap R'_j \neq \varnothing$ 且$|R'|<K$,则拆分方法如下:初始 $R'=R$,如果 R'中的两个元素 $R_i \cap R_j \neq \varnothing$就将 R_i,R_j 拆分为 $R_i-R_j,R_i \cap R_j,R_j-R_i$ 三个集合,替代原来的 R_i,R_j,这个操作不断进行,直到 R'中任何两个集合的交集都为空为止.当集合 R 或 R'符合上面的条件时, $avgdoc2$ 的最优值是

$$\min(avgdoc2)=\frac{\sum_{i=1}^{n}|R_i|}{n}.$$

这是 *avgdoc*2 最优值的上界.在这种划分中,每个查询的相关文档被划分在一个或几个集合中,在检索时,被选择的文档集合只含有与该查询相关的文档,没有与查询无关的文档.

5 算法复杂度分析

在类 Huffman 编码的划分算法中,主要的计算过程就是每次要找到需要合并的两个子集合的过程,实际上,求最优解的过程可以采用动态规划的算法:在第 1 次求待合并的集合时,需要计算两两子集合的合并代价,此时的比较次数为 $\frac{N(N-1)}{2}$;但当第 2 次求待合并子集合时,可以利用上次的部分计算结果,在上次的计算结果中,只有与上次合并的两个子集合有关系的结果是无效的,其他结果都可以继续利用,同时,需要添加此次新生成的集合与其余各集合的合并代价,此时的比较次数为 N–2 次;其后的合并计算以此类推,比较次数逐次递减.因此,对于模型 1,总的比较次数为

$$\frac{N(N-1)}{2}+(N-2)+(N-3)+...+1=\frac{N(N-1)}{2}+\frac{(N-2)(N-3)}{2}.$$

而对于模型 2,只要合并到结果个数为 K 时就可以结束.因此,比较的次数为

$$\frac{N(N-1)}{2}+(N-2)+(N-3)+...+K=\frac{N(N-1)}{2}+\frac{(N+K-2)(N-K-1)}{2}.$$

从而,模型 1 和模型 2 采用类 Huffman 编码的划分算法的时间复杂度是 $\theta(N^2)$.由于采用动态规划计算时需要保存前一次所有的比较结果,因此,所需要的空间开销为 $\frac{N(N-1)}{2}$.

6 划分模型作为文档集合划分的评价标准

文档集合划分作为分布式信息检索的一个步骤,对于它的评价可以采取分布式信息检索的结果作为评价的标准,分布式信息检索的评价通常采用检索评价的准确率与查全率来进行评价.但因为整个检索过程还涉及到“集合选择”、“单数据集合检索”以及“结果合并”等过程,这些部分都将对最终的检索结果产生影响,从而造成对文档集合划分评价的不准确.好的方法是对数据集合划分算法直接进行评价.

按照对文档集合划分问题的分析,模型 1 和模型 2 可以作为文档集合划分的评价标准.如果知道了查询的相关文档集就可以对不同的文档集合划分方案进行评价,确定哪一种划分方案更好.如果两种集合划分算法的划分子集个数不同,可以采用模型 1 来进行评价;如果两个集合划分算法的划分子集个数相同,则可以直接采用模型 2 进行评价.

具体的评价过程可以如下进行:如果要对某种文档集合划分算法 ξ 进行评价,首先采用算法 ξ 将文档集合划分为若干子集合,构建分布式信息检索系统,输入一定数量的查询,并在检索结果中找出查询所对应的全部相关文档.通过这个过程,我们就可以得到模型计算所需要的查询的相关文档集,从而可以将该算法划分的文档集合代入到模型 1 中,计算出对应的模型 1 的值.此时,模型 1 的取值就是该文档集合划分算法的评价值,同时,我们还可以利用本文给出的模型最优值求解算法,计算出最优的模型取值作为评价参考.

上面的评价方法是基于给定查询测试集的,测试集的客观性对文档集合划分方案的评价将产生很大的影响.一个“好”的文档集合的划分结果,只是在给定查询集下的“好”的划分结果,对于其他查询集则不一定是好的划分结果.在实际中,可以采用查询日志对划分的结果进行评价,查询日志具有很好的预测未来查询的能力.查询在日志中重复出现的情况是非常高的,在日志中出现的查询,在未来可能会以一定的概率重复出现,从而利用查询日志可以预测未来查询的出现情况.这样,如果我们用查询日志来评价划分的结果,那么对于未来的查询也有很好的指导意义.

7 实 验

实验的目的一方面是通过实验对文档划分模型的最优解进行分析与讨论,另一方面是验证本文提出的文档划分模型可以作为有效的文档集合划分评价标准,可以对不同的文档集合划分算法进行评价比较.

7.1 实验的设计

7.1.1 文档划分模型最优解的分析实验

在这个实验中,我们通过给定一个文档集合,并给出一个查询集对应的相关文档集合,讨论在给定的相关文档集合的条件下,文档划分模型的最优解.

7.1.2 文档划分模型作为文档划分评价标准的实验

在这个实验中,对于一个给定的文档集合,我们分别采用两种不同的文档集合划分方法对文档集合进行划分:一种方法是基于内容聚类的方法;另一种方法是随机的文档划分方法.基于内容的聚类方法采用了 *K*-Means 聚类算法,文档相似度的计算采用了向量空间模型,将文档聚成 52 个子集合(52 的选择是由实验 1 中模型 1 最优解得到的);对于随机的文档划分方法,则是随机地将文档划分到 52 个子集合中.由于基于内容聚类的方法考虑了文档的相关性,因此,应该比随机的方法在文档划分中更加有效,这在文献[2]中也有证明.我们通过对比这两种方法计算的模型 2 的值,来观察模型 2 是否能对这两种划分方法进行很好的评价.

7.2 实验数据集合的构建

实验采用了 TREC 中 Web Track 的查询集作为数据集,将 2002 年~2004 年的查询集合并起来,这些查询的相关文档构成了整个文档集,3 年总的查询个数为 324 个,相关集合中文档个数为 3 853 个,其中,不同文档个数为 3 778 篇,有 69 篇文档同时是 2 个查询的相关文档,有 3 篇文档同时是 3 个查询的相关文档.

7.3 实验结果及分析

在文档划分模型最优解实验中,我们对上面的文档集合进行划分,利用本文提出的模型最优解求解算法,求出模型 1 和模型 2 的最优解,实验结果如图 3 所示.

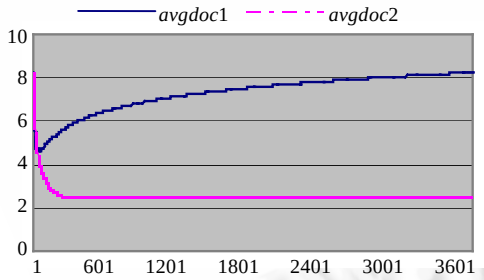

Fig.3 The model value in different partition parts

图 3 不同子集合个数下模型值变化曲线

图 3 是划分的子集合个数从 1 个变化到 3 778 个时,模型 1 的 *avgdoc*1 的取值与模型 2 最优值 *avgdoc*2 变化的曲线,其中横坐标是划分的子集合个数,纵坐标是模型 1 *avgdoc*1 和模型 2 *avgdoc*2 值的对数.

正如前面分析的那样,从图 3 中可以看出,模型 2 的最优值随着集合个数的增加逐渐下降,当达到某个临界点时不再下降.根据模型 2 的分析,此时的最优值是 $\frac{\sum_{i=1}^{n}|R_i|}{n}$,对于本测试集,其值为 11.89(3853/324),从 *avgdoc*1 的计算结果看,模型 1 在划分的子集合个数为 52 时取得最优值,此时,*avgdoc*1 的值为 102.04,也就是说,按照这种划分算法,对于规模是 3 778 篇文档的文档集合,查询这 324 个查询平均需要查询的文档数为 102 篇,这样,对于每个查询只要检索相当于原文档集合的 2.6%,就可以取得和检索全部文档集一样的检索结果,需要检索的文档数量大为减少.

进一步地,在文档划分模型作为文档划分评价标准的实验中,分别将基于内容聚类的划分方法与随机的划分方法进行了比较,同时与最优划分方法也进行了比较.如果模型 1 和模型 2 是一种有效的评价方法,那么期望得到的结果是,基于内容聚类的划分方法要好于基于随机划分的结果,3 种划分的评价结果见表 1.

从表 1 可以看出,基于内容聚类算法的 *avgdoc*1 和 *avgdoc*2 的值都小于随机划分的结果,这说明,基于内容的划分方法“平均待查文档数”较小.也就是说,对于每个查询,按照基于内容聚类的方法,平均只需检索 274 篇文

档就可以检索到该查询的全部相关文档;而按照随机划分的方法,则平均需要检索 602 篇文档才能检索到该查询的全部相关文档.根据以往的研究,基于内容的聚类方法应该好于随机划分的结果,而模型 1 和模型 2 的值很好地体现了这一点,这说明,模型 1 和模型 2 可以作为文档集合划分的有效评价方法.另外,通过与最优划分策略比较可以看出,基于内容聚类的划分策略与最优划分方案之间还存在一定的差距.

Table 1 Comparison of three partition methods

表 1 3 种划分算法结果比较

	Optimization	Content-Based	Random
*avgdoc*1	102.04	325.59	653.15
*avgdoc*2	50.04	273.59	601.15

8 结束语

文档集合划分问题是分布式信息检索中的一个重要问题.本文对分布式信息检索的文档划分问题作了比较深入的分析,给出文档划分的问题定义,构建了两个模型来刻画文档集合的划分问题,提出了一个类 Huffman 编码的最优化解法,并将这两个模型作为文档集合划分算法的评价指标.无论是理论上的分析,还是实验的结果都证明,本文提出的两个文档划分模型可以作为有效的文档集合划分的评价方法.利用这两个模型可以直接对文档集合划分方法进行评价,而不需要通过分布式信息检索的结果来间接地评价文档集合的划分算法,从而克服了“集合选择”、“单数据集合检索”以及“结果合并”等过程对于评价的影响.因此可以说,本文提出的两个模型是一种有效的分布式信息检索的文档划分评价方法.

References:

[1] Croft WB. Advances in Informational Retrieval. Norwell: Kluwer Academic Publishes, 2000. 127−150.

[2] Xu JX, Croft WB. Cluster-Based language models for distributed retrieval. In: Gey F, ed. Proc. of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 1999. 254−261.

[3] Callan JP, Lu ZH, Croft WB. Searching distributed collections with inference networks. In: Fox EA, ed. Proc. of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 1995. 21−28.

[4] French JC, Powell AL, Viles CL, Emmitt T, Prey KJ. Evaluating database selection techniques: A testbed and experiment. In: Croft WB, ed. Proc. of the ACM Special Interest Group on Information Retrieval Conf. New York: ACM Press, 1998. 121−129.

[5] Sogrine M, Patel A. Evaluating database selection algorithms for distributed search. In: Lamont GB, ed. Proc. of the 2003 ACM Symp. on Applied Computing. New York: ACM Press, 2003. 817−822.

[6] Luo S, Callan J. Unified utility maximization framework for resource selection. In: Grossman D, ed. Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2004. 32−41.

张刚(1977－),男,黑龙江牡丹江人,助理研究员,主要研究领域为信息检索,自然语言处理.

谭建龙(1974－),男,博士,副研究员,主要研究领域为数据流,大规模字符串匹配技术.