

基于文档合并的 XML 无线数据广播调度算法

吴晶晶, 毛鼎鼎, 朱 良, 张 璞, 孙未未

(复旦大学计算机科学技术学院, 上海 201203)

摘 要: 在可扩展标记语言(XML)无线数据广播中, 数据以 XML 文档为基本单位进行广播, 然而 XML 文档间的冗余信息会降低带宽资源的利用率。为解决该问题, 提出一种有效的调度算法, 分析文档合并对数据广播的性能影响, 得出文档间亲密度的衡量标准, 并将亲密度高文档进行合并以减少冗余信息。实验结果证明, 该算法可以提高无线数据广播性能, 节约带宽资源。

关键词: 可扩展标记语言; 数据广播调度; 文档合并; 调谐时间; 访问时间

XML Wireless Data Broadcast Schedule Algorithm Based on Document Mergence

WU Jing-jing, MAO Ding-ding, ZHU Liang, ZHANG Pu, SUN Wei-wei

(School of Computer Science, Fudan University, Shanghai 201203, China)

【Abstract】 An eXtensive Markup Language(XML) document is the broadcast unit in XML data broadcast environment. Data broadcast is an efficient way for information dissemination in wireless mobile environments. The efficiency of broadcast is reduced by the redundant information between XML documents. In order to solve the problem, this paper proposes an efficient schedule algorithm. An affinity of XML documents is derived by analyzing the effect of merging documents. Redundant information is reduced by recursively merging the documents with the largest affinity. Experimental result shows that this algorithm can enhance the effective of broadcast and save the limited bandwidth.

【Key words】 eXtensive Markup Language(XML); data broadcast schedule; document mergence; tuning time; access time

DOI: 10.3969/j.issn.1000-3428.2011.14.008

1 概述

无线广播是无线环境中的基本数据访问模式之一, 具有节能、提高带宽利用率等特点。可扩展标记语言(eXtensive Markup Language, XML)作为数据交换的标准越来越广泛地应用到无线环境中。因此, XML 数据广播已成为一个新的研究热点。已有传统数据广播调度工作遵循广播数据项的原子性假设, 即数据项是不可分割的数据单元, 也是用户下载的最小数据单位。XML 是半结构化数据, 根据查询语义的不同, 用户的查询结果可能是一个 XML 文档, 也可能是 XML 文档中的部分分支。基于 XML 文档的非原子性和冗余性, 本文提出一种有效的调度算法提高广播性能。

2 相关工作和背景知识

2.1 XML 数据广播调度技术

文献[1]研究了 XML 索引技术。在 XML 数据广播调度技术方面, 文献[2]提出一种按需环境下的文档分割算法, 根据服务器端用户的查询情况, 将每个文档分割为 2 个部分。文献[3]提出一种更高效的基于查询分组的文档分割算法。文献[2-3]主要从用户角度出发, 尽量减少用户下载冗余数据。

2.2 背景知识

衡量无线数据广播系统性能的指标主要有^[4]:

(1)访问时间(AT)指从用户提出请求开始到用户从信道下载完所需的全部结果的时间。AT 决定了用户的等待时间。

(2)调谐时间(TT)指在完成一个访问请求期间, 用户保持侦听信道的总时间。在不监听信道的时间里, 用户可以进入休眠状态节约能量。因此, TT 决定用户的能量消耗情况。在 XML 数据广播中, 同样采用 AT 和 TT 作为广播性能的衡量

标准。

3 基于文档合并的调度算法

在周期广播模式下, 假定移动用户的请求分布模式是已知并且稳定的。服务器根据这些信息进行广播调度, 在生成了相应的广播调度序列后不断地循环广播出去。目前, 传统数据广播多数据项调度方法研究都集中在这种模式^[5]。本文也采用这种模式。

由上文可知, 文档分割技术减少了用户下载的冗余信息, 但同时增加了广播数据的大小。由于用户提交的查询为 XPath 模式, 一般有多个文档与查询相匹配, 因此若将同时被用户请求率高的文档进行合并, 则可以减少一部分用户的 TT, 同时减少广播数据。

本文基于文档合并的调度算法的主要思想是: 计算每 2 个文档间的关系密切程度, 通过不断将密切度最高的文档合并以达到减少广播数据总长度的目的。

本文算法有 2 个关键性问题:

- (1)如何定义 2 个文档间的关系密切程度;
- (2)如何定义算法的结束条件。

通过分析文档合并对广播性能指标 AT 和 TT 的影响, 解决上述 2 个问题。符号说明如表 1 所示。

基金项目: 国家自然科学基金资助项目“XML 格式数据的无线广播技术研究”(61073001)

作者简介: 吴晶晶(1986—), 女, 硕士研究生, 主研方向: XML 无线数据广播; 毛鼎鼎、朱 良, 硕士研究生; 张 璞, 本科生; 孙未未, 副教授、博士

收稿日期: 2011-02-21 **E-mail:** wwsun@fudan.edu.cn

表1 符号说明

符号	说明
d_i	一个 XML 文档
D	服务器端的整个文档集合 $D=\{d_1, d_2, \dots, d_n\}$
q_i	用户提出的一个 XPath 查询
Q	所有查询的集合 $Q=\{q_1, q_2, \dots, q_m\}$
$QDS(q_i)$	q_i 需要访问的 XML 文档集合
$freq(q_i)$	查询 q_i 的访问频率

3.1 基本概念

假设 XML 文档合并后新的文档集合为 $\{d_1, d_2, \dots, d_n\}$, 若将 XML 文档作为数据项, XML 查询作为多数据项请求, 则可以采用传统的多数据项调度策略对文档集进行排序调度。

文献[5]分析传统数据广播中多数据项查询的 AT 。设 σ 为广播序列, $\{d_1, d_2, \dots, d_m\}$ 为查询 q_i 请求的数据集合, δ_i 为数据 d_i 、 d_{i+1} 之间的间隔, 那么查询 q_i 的平均访问时间为:

$$AT^{avg}(q_i, \sigma) = BSize - \frac{1}{2BSize} \sum_{j=1}^m \delta_j^2 \quad (1)$$

其中, $BSize$ 为广播序列的总长度。总访问时间为:

$$TAT = \sum_{q_k \in Q} AT^{avg}(q_k, \sigma) \times freq(q_k) \quad (2)$$

查询的调谐时间为其请求的全部数据的大小之和, 与数据的调度顺序无关。因此, 查询的调谐时间为:

$$TT(q_i) = \sum_{i=1}^m Size_i \quad (3)$$

其中, $Size_i$ 文档 d_i 的长度。总调谐时间为:

$$TTT = \sum_{q_k \in Q} (TT(q_k) \times freq(q_k)) \quad (4)$$

在上述公式基础上, 分析文档合并对 TAT 和 TTT 的影响。

3.2 文档合并对广播性能的影响分析

由于采用传统多数据项广播的调度策略对 XML 文档进行调度, 广播序列长度 $BSize$ 等于文档总长度 L 。根据式(1)平均访问时间随 $BSize$ 的减小而减小。假设文档集合总长度为 L , 平均 2 个文档合并后的长度减少率为 λ , 合并次数为 x , 则合并后的文档集合总长度为:

$$L' = L - \frac{2L\lambda x}{|D|} \quad (5)$$

在式(5)中, 合并后的文档总长度随合并次数 x 的增加而减少, 因此, 增加合并次数可以减少平均访问时间。

假设查询集合为 Q , 对于任意的文档 d_i 、 d_j , 设同时请求 d_i 、 d_j 的查询集合为 $Q_{i \cap j}$, 只请求 d_i 的查询集合为 Q_i , 只请求 d_j 的查询集合为 Q_j , 则合并文档 d_i 、 d_j 后的调谐时间为:

$$\Delta TT = (|Q_i| + |Q_j|) \times \frac{L(1-2\lambda)}{|D|} - |Q_{i \cap j}| \times \frac{2L\lambda}{|D|} \quad (6)$$

结合式(6)合并后的总调谐时间为:

$$TTT' = TTT + \Delta TT \quad (7)$$

从式(7)可以看出, 当 $\Delta TT < 0$ 时, 总调谐时间将减少。根据式(6)发现 ΔTT 随 λ 和 $|Q_{i \cap j}|/(|Q_i| + |Q_j|)$ 的增大而减少。实际上, λ 会随合并次数的增加而减少, 这是因为文档间的重叠信息是有限的, 合并过多的文档会降低合并效率。因此, 选取 $|Q_{i \cap j}|/(|Q_i| + |Q_j|)$ 大的文档进行合并, ΔTT 可能取到较小值。所以, 本文算法将 $|Q_{i \cap j}|/(|Q_i| + |Q_j|)$ 作为评估文档 d_i 、 d_j 间亲密度的标准。用 TTT 作为结束条件, 在 TTT 不增加的前提下, 合并文档, 减少 TAT 。

3.3 文档间关系亲密度

本文用文档间的关系密切度评估任意 2 个文档间的可合并程度。

定义 1 文档 d_i 、 $d_j (i \neq j)$ 间的关系密切度如式(8)所示:

$$aff(d_i, d_j) = \frac{\sum_{q_k \in Q} I_k(i, j) \times freq(q_k)}{1 + \sum_{q_k \in Q} (U_k(i, j) - I_k(i, j)) \times freq(q_k)} \quad (8)$$

当 $i=j$ 时, $aff(d_i, d_j)=0$ 。其中, 若查询 q_k 同时访问 d_i , 则 $d_j \cap_k(i, j)=0$, 否则为 0。同理, 若查询 q_k 同时访问 d_i , 则 $d_j \cup_k(i, j)=2$, 若查询 q_k 只请求其中一个, 则 $U_k(i, j)=1$, 不请求为 0。

根据上文分析, 文档间的关系密切度可以理解为: 同时请求文档 d_i 、 d_j 的查询之和与只请求文档 d_i 或只请求文档 d_j 的查询之和的比值。 $aff(d_i, d_j)$ 越大, 合并文档 d_i 、 d_j 后的 ΔTT 越小。通过实验可以验证, TTT 随着合并次数的增加, 有先下降后上升的趋势, 本文中的算法是在 TTT 小于合并前的这段区间内, 合并文档从而减少 TAT 。

定义 2 假设 n 为文档总数量, 则亲密度矩阵 AM 为一个 $N \times N$ 的矩阵, 且矩阵中的元素为:

$$AM[i][j] = aff(d_i, d_j)$$

3.4 算法描述

在对亲密度矩阵初始化后, 本文算法开始根据文档间的关系密切度合并文档。

算法 文档合并

输入 文档集合 D , 查询集合 Q

输出 合并后的文档集合

(1) 根据 D 和 Q 计算 AM , T 为合并前初始化的 TTT , T' 标记每次合并后的 TTT 。

(2) 根据 AM 找出关系密切度最大的 2 个文档 d_i 、 d_j 。

(3) 使 T' 等于合并文档 d_i 、 d_j 后的 TTT , 若 $T' > T$, 则跳转至第(6)步。

(4) 合并文档 d_i 、 d_j 。

(5) 更新 AM , 查询集合, 文档集合。跳转至第(2)步。

(6) 算法结束。

下面通过一个具体实例解释该算法。表 2 和表 3 分别描述了文档集合和查询集合的情况。

表2 文档集合及其长度

文档集合	长度
d_1	3
d_2	5
d_3	5
d_4	10

表3 请求查询集合及其访问频率

Q	请求文档集合	访问频率
q_1	d_1, d_2	2
q_2	d_2, d_3	3
q_3	d_1, d_2, d_4	3
q_4	d_3, d_4	1

根据查询请求文档集合及其访问频率可以计算出亲密度矩阵 AM 。例如:

$$AM[2][3] = aff(d_2, d_3) = 0.43$$

以此类推, 最终初始化的亲密度矩阵如表 4 所示。文档合并的过程如表 5 所示。其中, 第 0 步为初始化阶段。第 3 步中由于合并文档 d_1 、 d_4 后的 TTT 大于初始化时的 TTT , 因此算法结束, 最终的文档合并结果集合为 $\{d_1, d_4\}$ 。

表4 初始化的亲密度矩阵

文档	d_1	d_2	d_3	d_4
d_1	0.00	1.25	0.00	0.75
d_2	1.25	0.00	0.43	0.43
d_3	0.00	0.43	0.00	0.14
d_4	0.75	0.43	0.14	0.00

表5 文档合并过程

序号	合并文档	TTT	L
0	-	115.000 0	23.00
1	d_1, d_2	108.000 0	21.00
2	d_1, d_3	114.250 0	18.25
3	d_1, d_4	123.187 5	-

在通过基于文档合并的调度算法合并 XML 文档后, 会产生一个新的 XML 文档集合, 本文采用文献[5]中的调度算法对该文档集合进行调度。

3.5 算法性能分析

假设文档集合长度为 n , 查询集合长度为 m 。在算法描述的第(5)步更新 AM 时, 发现受到合并影响的元素为 $AM[d_i][d_j]$ 和 $AM[d_i][d_l]$, 其中, $d_i \neq d_j \neq d_l$ 。因此, 更新 AM 的复杂度为 $O(mn)$ 。在最差的情况下, 合并次数为 $n-1$, 算法复杂度是 $O(mn^2)$ 。

4 实验及结果分析

4.1 实验环境及参数设置

本文主要比较合并前与合并后的用户的 AT , 采用文献[5]中的调度算法作为基本的调度策略, 随机生成查询的访问频率。在实验中, 发现压缩合并后的 XML 文档可以进一步减少文档间冗余信息, 因此, 将压缩后的 zip 文件作为广播数据。

实验采用 Java1.6.0 编译环境, 在 WinXP 的平台上进行。实验对象是 NITF(News Industry Text Format) DTD 所定义的 XML 文档, 由 IBM 的 XMLGenerator 生成。采用 YFilter 中的 XPath 生成器生成用户的查询。默认查询的最大路径为 10, “/” 和 “*” 的概率为 10%。在本实验中, 将用户下载的数据的字节数代表 AT 和 TT 。变化的实验参数设置见表 6。

表6 实验参数设置

参数	取值范围	默认值	说明
查询类型数	50~400	200	查询类型的数量
文档数	500~3 000	1 000	服务器端 XML 文件的总数

4.2 实验结果

本文采用平均调谐时间($avgTT$)作为衡量 TT 的标准, $avgTT$ 等于 TTT 与查询总数的比值。图 1 展示了在查询类型数量和文档数量取默认值时, $avgTT$ 随合并次数的变化情况。合并总数为 255 次。可以看出, $avgTT$ 在这 255 次合并中先减少后增加, 减少的原因是通过合并关系密切度高的文档, 使得 TTT 减少; 增加的原因是过多的合并导致查询下载更多冗余信息。

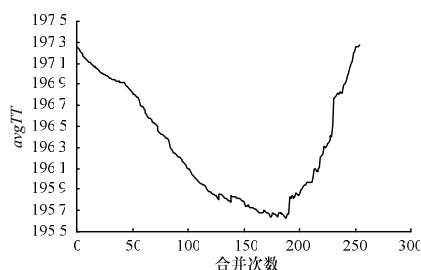
图1 合并次数对 TT 的影响

图 2 展示了查询类型对 AT 的影响。本文采用平均访问时间($avgAT$)作为衡量 AT 的标准, $avgAT$ 等于 TAT 与查询总数的比值。可以看出, $avgAT$ 随着查询类型的增加而增加,

经过基于文档合并的调度算法后, 减少了文档间的冗余信息, AT 得到了改善。当查询种类为 50 时, 本文算法减少了 16.9% 的 $avgAT$ 。

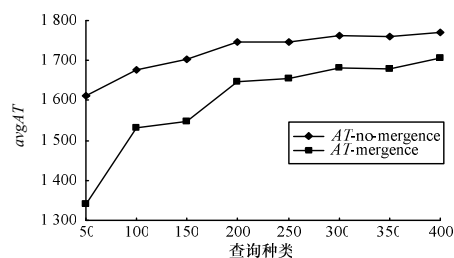
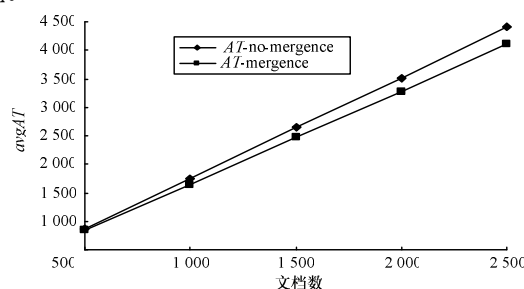
图2 查询种类对 AT 的影响

图 3 展示了文档数量对 AT 的影响。 $avgAT$ 随着文档数量的增加而增加, 这是因为与 XPath 查询匹配的文档数量增加了。而本文调度算法随着文档数量的增加优化效果更明显, 减少了更多的文档间冗余信息, 说明本文算法具有良好的适应性。

图3 文档数对 AT 的影响

5 结束语

本文提出一种基于文档合并的 XML 数据广播调度算法。通过分析文档合并对广播性能的影响, 定义一种合理的文档间关系密切度的衡量标准。实验证明本文算法在不影响 TT 的基础上, 可以有效减少文档间的冗余信息, 从而减少 AT 。

参考文献

- [1] 杨进才, 龚 松, 许沛华, 等. 移动环境中自适应的 XML 广播索引[J]. 计算机工程, 2009, 35(13): 46-48, 74.
- [2] Qin Yongrui, Sun Weiwei, Zhang Zhuoyao. An Efficient Document-split Algorithm for On-demand XML Data Broadcast Scheduling[C]//Proceedings of the 2nd IET International Communication Conference on Wireless Mobile & Sensor Networks. Shanghai, China: [s. n.], 2007.
- [3] Qin Yongrui, Sun Weiwei, Zhang Zhuoyao, et al. Query-grouping Based Scheduling Algorithm for On-demand XML Data Broadcast[C]//Proceedings of the 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computing. Dalian, China: [s. n.], 2008.
- [4] Xu J, Lee D, Hu Q, et al. Data Broadcast[C]//Proceedings of Handbook of Wireless Networks and Mobile Computing. [S. l.]: John Wiley & Sons, 2002.
- [5] Chung Yon-Dohn, Kim Myoung-Ho. A Wireless Data Clustering Method for Multipoint Queries[J]. Decision Support Systems, 2001, 30(4): 469-482.

编辑 陆燕菲