



异构文本数据转换中 XML 解析方法对比研究

何卓桁¹, 刘志勇¹, 李 璐², 李长明³, 张 琳⁴

(1. 东北师范大学 信息科学与技术学院, 长春 130024; 2. 同济大学 软件学院, 上海 200092;
3. 长春光华学院 电气信息学院, 长春 130033; 4. 吉林大学 软件学院, 长春 130012)

摘 要: 对异构文本数据转换过程中解析 XML 文本的 DOM、SAX、JDOM、DOM4J 方法进行对比研究, 以解析时间、内存堆占用空间、CPU 占用率为评价指标来判定 4 种解析方法的优劣。该评价方法的优势在于当数据量或数据属性发生变化时, 4 种解析方法对评价结果的影响仍具有良好的区分度。通过对 10 份 Web 日志异构文本数据转换后的 XML 数据集进行比较, 实验结果表明, 当数据量增大且以解析时间为重点时, DOM4J 解析方法优于其他 3 种解析方法, 当以空间占用为重点时, SAX 解析方法优于其他 3 种解析方法。

关键词: 异构文本; XML 解析; 数据结构转换; 时间复杂度; 空间复杂度

开放科学(资源服务)标志码(OSID):



中文引用格式: 何卓桁, 刘志勇, 李璐, 等. 异构文本数据转换中 XML 解析方法对比研究[J]. 计算机工程, 2020, 46(7): 286-293, 299.

英文引用格式: HE Zhuoheng, LIU Zhiyong, LI Lu, et al. Comparative study of XML parsing methods in heterogeneous text data conversion[J]. Computer Engineering, 2020, 46(7): 286-293, 299.

Comparative Study of XML Parsing Methods in Heterogeneous Text Data Conversion

HE Zhuoheng¹, LIU Zhiyong¹, LI Lu², LI Changming³, ZHANG Lin⁴

(1. School of Information Science and Technology, Northeast Normal University, Changchun 130024, China;
2. School of Software, Tongji University, Shanghai 200092, China; 3. School of Electrical and Information Engineering, Changchun Guanghua University, Changchun 130033, China; 4. School of Software, Jilin University, Changchun 130012, China)

[Abstract] This paper compares and studies the DOM, SAX, JDOM, DOM4J methods for parsing XML texts in heterogeneous text data conversion. The pros and cons of the four parsing methods are judged based on parsing time, memory heap space, and CPU occupancy rate. The advantage of this evaluation method is that when the amount of data or data attributes change, the impact of the four analytical methods on the evaluation results still has a good degree of discrimination. By comparing 10 converted XML datasets of heterogeneous text data of Web log, experimental results show that when the amount of data increases and the analysis time is mainly concerned, the DOM4J parsing method is superior to the other three analysis methods. When space occupation is mainly concerned, the SAX parsing method is superior to the other three analysis methods.

[Key words] heterogeneous text; XML parsing; data structure conversion; time complexity; space complexity

DOI: 10.19678/j.issn.1000-3428.0054925

0 概述

由于异构文本数据具有数据量大、形式多样且来源复杂等特点, 在数据预处理工作中, 存在查找有

效信息困难的问题。为了对数据进行过滤并达到筛选有效信息的目的, 需要对数据结构进行转换, 保证数据的统一化, 从而简化后续文本的处理工作。在异构文本的数据预处理工作中, 异构数据的转换是

基金项目: 吉林省教育厅“十三五”科学技术研究规划项目“基于高校学生综合素质测评数据预测职业发展方向研究”(202118628); 吉林省教育厅新工科研究与实践项目“U-G-E‘卓越软件工程师’人才培养模式与实践教学深化改革”(131003229)。

作者简介: 何卓桁(1994—), 男, 硕士研究生, 主研方向为数据挖掘、数据预处理; 刘志勇(通信作者), 副教授、博士; 李 璐, 本科生; 李长明, 硕士; 张 琳, 研究员。

收稿日期: 2019-05-15 **修回日期:** 2019-07-22 **E-mail:** lzy600@qq.com

不可或缺的步骤,主要分为直接转换和间接转换。直接转换是利用正则表达式对异构文本进行过滤并建立其对应结构,间接转换是将异构文本转换为半结构化的XML文本,以XML文本为桥梁转换为结构化文本^[1]。间接转换已经得到了学者们的普遍认可,其转换过程主要包括2个阶段,第1阶段是通过制定转换规则将异构文本转换为XML文本,第2阶段是通过一定的解析方法将XML文本转换为结构化文本。采用XML文本作为中间转换的标准,优势在于利用XML文本的分层嵌套格式,以及在分层表示的各个元素中均包含属性和值^[2],使得语义表达能力突出。另外,XML具有格式规整的特点,XML文档不需要符合特定文档类型定义(Document Type Definition, DTD)或者架构^[3],这些特点使得XML文档表达的Web内容能够更好地被用户理解。因此,XML适合存储半结构化的数据。

第1阶段中转换方法的研究相对比较成熟,主要是对超文本标记语言进行去标记、分类,制定XML模板,主要研究工作包括:文献[4]利用DTD或者由一种用于描述和规范XML文档逻辑结构的语言(Schema)制定对应的规则,生成XML Schema。文献[5]通过从XML文件中提取结构信息来创建一个临时的DTD,将XML文件映射为对象数据库。文献[6]通过分析DTD和XML Schema 2种模式的不同之处,参考基于DTD的XML函数依赖的相关研究,提出XML Schema形式化定义和XML的轴元素定义,给出基于XML Schema标准的XML函数依赖定义以及其推理规则集,规范了XML文档。文献[7]通过模板建立Schema并将XML文本结构存入其中后再进行解析。文献[8]对DTD进行深入研究,并讨论XML架构受元素声明一致性(Element Declaration Consistency, EDC)规则的影响。文献[9]针对Web日志中的元素存在属性和值,利用XML文本自身的分层结构与其相关联的优势,使得Web日志内容能够更好地被表达。第2阶段的解析方法主要有DOM、SAX、JDOM和DOM4J等4种,然而这4种方法在什么情况下才会更加有效,目前还缺乏科学的实验论证,尚未形成统一的结论。

综上所述,XML文本在异构文本数据转换过程中起到了至关重要的作用,重点关注在异构数据转换过程的第2阶段,即将XML文本通过解析方法转换为结构化数据。本文结合方法组合的思想,采用多组实验进行比较^[10],在同一实验环境下,保证限定条件的统一,以多角度的方式对DOM、SAX、JDOM和DOM4J解析方法进行研究。相较于以往的对比较研究^[11],本文引入内存占用空间、CPU占用率和解析时间作为评价指标,相比单一

的以效率为评价指标的方法,本文的评价指标方法更为全面、客观和准确。以加权的方式考虑不同指标的影响因素,使解析方法之间的区分度增大,结果更加直观。同时,在统一评价指标上设置多组实验,对4种解析方法在不同的价值取向,验证其解析方法的优劣性,为后续研究者针对不同的研究目的提供更加直观切实的研究方法。

1 解析方法

以XML为介质,采用4种目前主流的解析方法DOM、SAX、JDOM和DOM4J对数据进行转换。其中,DOM是目前解析XML文本的基础解析方法,通过树形结构存储,该方法与XML存储方法吻合,使用户能够更好地理解。SAX采用流处理方式,占用内存少,适用于处理文本量大的工作。JDOM结合了DOM与SAX的优点,基于树形结构,提供更加简单的逻辑访问方法。DOM4J是JDOM的一种智能分支,它合并了包括集成XPath支持、XML Schema支持以及用于长文档或流化式文档的基于事件处理的功能。

1.1 DOM解析方法

DOM解析方法是将XML文本转换为对象模型,运用树结构对信息进行存储,通过接口的方式进行访问且可以访问任意节点^[12]。关于DOM解析方法的研究,文献[13]利用DOM树中的文本内容和层次结构对Web中的菜单和导航指示器的关键信息进行提取,通过在目标网页中点选元素的方式,自动生成基于DOM路径的抽取模板,从而达到解析并提取信息的目的。文献[14]针对本地存储结构化数据的XML文档,设计一个基于DOM树的轻量级文档解析库。但上述方法也存在一定的缺陷,如将文本转换为树结构时,若需要转换的文件较大,则对树结构的遍历十分耗时,将导致整个解析过程十分缓慢。

1.2 SAX解析方法

SAX解析方法是以时间为驱动的API,采用类似流处理的方式在进行扫描文本的同时自顶向下依次完成解析任务^[15]。在解析XML过程中,SAX占用的内存少且速度快。关于SAX解析方法的研究,文献[16]提出基于SAX解析过程,利用列表以及关系指针2种方法相接合的方式来处理XPath查询的QXSLList方法,通过层次值计算判断节点的结构关系,利用关系指针链接多个候选节点列表来获取查询结果。虽然SAX在处理关系层次较多以及文本数据量大的情况下表现优异,但是也存在SAX无法随机访问XML节点,也无法对XML文本进行修改,只能对文本进行读取任务的缺点。

1.3 JDOM 解析方法

JDOM 解析方法是基于树形结构,内置 Xerces 解析器,使用具体类的文档解析模型,运用树结构对信息进行存储,它所包含的转换器将 JDOM 表示输出成 SAX 事件流、DOM 模型^[17]。关于 JDOM 解析方法的研究,文献[18]利用其解析 XML 和 Schema 文件,完成了异构数据的转换,该解析方法简化了异构数据转换的流程,并且保证了关系数据信息的完整性,但是这需要用户充分地理解 XML 文本,说明 JDOM 解析方法缺乏一定的灵活性。

1.4 DOM4J 解析方法

DOM4J 解析方法是 JDOM 的分支,集成了 DOM 和 SAX 的 XML 文件解析器,提供大量接口用于对 XML 文件进行处理,且 DOM4J API 和标准 DOM 接口具有并行访问功能^[19]。关于 DOM4J 的研究,文献[20]以 XML 文本作为数据库存储学生信息,利用 DOM4J 树结构进行解析,结果表明,其解析时间较 DOM 解析方法短,这说明 DOM4J 在理论上较 DOM 表现良好。

通过对大量文献进行分析,结果发现,间接转换过程中的主要问题在于对目前主流的 4 种解析方法的优劣存在矛盾的结论。在各自领域中,运用不同的解析方法均能达到其预期效果,并不能体现出解析方法的优劣。因此需要在相同条件下进行比较并通过实验证明目前主流的 4 种解析方法的优劣。

2 算法分析

在相同条件下,实验利用 4 种解析方法对相同的数据集进行解析,其解析机制是根据文档内容对其节点和元素进行读取并输出,解析的质量以消耗的时间和空间为评价指标,处理结果的正确性取决于程序是否能够运行至结束。在实验过程中,数据集的数据量和属性个数在理论上对实验结果的精度没有影响,在算法分析过程中主要以解析方法的时间开销来对算法的优劣性进行区分。

4 种解析方法的时间开销是实验中对文档进行解析所需要的时间,在本节中主要针对解析文档时处理耗费的运行时间复杂度进行理论计算。其中,对代码的运行次数记为 n ,每行代码的时间开销记为 Cx , x 代表程序对应的行,以下是 4 种解析方法的运行时间复杂度的分析。

DOM 解析方法的运行时间复杂度如表 1 所示,由此可知,由于 $C8$ 远远大于 $C1 \sim C7$,因此 $T(n) = C8n^3 = O(n^3)$ 。

表 1 DOM 解析方法的运行时间复杂度

Table 1 Running time complexity of DOM parsing method

程序段	时间 开销	次数
for(int i=0;i<list.getLength();i++) {	C1	n
NodeRemote_ip = list.item(i);	C2	n
NodeListRemote_ipList = Remote_ip.getChildNodes();	C3	n
for(int j=0;j<Remote_ipList.getLength();j++) {	C4	n^2
Node info = Remote_ipList.item(j);	C5	n^2
NodeList attribute = info.getChildNodes();	C6	n^2
for(int k=0;k<attribute.getLength();k++) {	C7	n^3
if(attribute.item(k).getNodeName() != "#text") {	C8	n^3

SAX 解析方法的运行时间复杂度如表 2 所示,由此可知,由于 $C3$ 远大于 $C1 \sim C2$,因此 $T(n) = C3n = O(n)$ 。

表 2 SAX 解析方法的运行时间复杂度

Table 2 Running time complexity of SAX parsing method

程序段	时间 开销	次数
SAXParserFactorysFactory = SAXParserFactory.newInstance();	C1	n
try {SAXParsersaxParser = sFactory.newSAXParser();	C2	n
SAXParseHandlersaxParseHandler = new SAXParseHandler();	C3	n

JDOM 解析方法的运行时间复杂度如表 3 所示,由表 3 可知,由于 $C6$ 远大于 $C1 \sim C5$,因此 $T(n) = C6n^2 = O(n^2)$ 。

表 3 JDOM 解析方法的运行时间复杂度

Table 3 Running time complexity of JDOM parsing method

程序段	时间 开销	次数
for(int i=0;i<logList.size();i++) {	C1	n
Element log = (Element) logList.get(i);	C2	n
List <Element> content = log.getChildren();	C3	n
for(int j=0;j<content.size();j++) {	C4	n^2
System.out.println(((Element) content.get(j)).getName() + " : " + ((Element)	C5	n^2
content.get(j)).getValue());	C6	n^2

DOM4J 解析方法的运行时间复杂度如表 4 所示,由此可知,由于 $C6$ 远大于 $C1 \sim C5$,因此 $T(n) = C6n^2 = O(n^2)$ 。

表 4 DOM4J 解析方法的运行时间复杂度

Table 4 Running time complexity of DOM4J parsing method

程序段	时间 开销	次数
for(Iterator i = list.elementIterator(); i.hasNext();){	C1	n
Element log = (Element) i.next();	C2	n
for(Iterator j = log.elementIterator(); j.hasNext();){	C3	n^2
j.next().getName() + " : " + ((Element) j.next().getText());	C4	n^2
Element node = (Element) j.next();	C5	n^2
System.out.println(node.getName() + " : " + node.getText());	C6	n^2

根据以上 4 种解析方法运行时间复杂度的分析可知, SAX 解析方法的运行时间复杂度相比其他 3 种解析方法都较小, 且 DOM 解析方法的运行时间复杂度最大。这是因为 SAX 解析方法采用的是流式处理文件的方法, 即用即停, 不需要将 XML 文本存入内存中, 理论上适合用于数据量较大的情况。DOM、JDOM 以及 DOM4J 解析方法均需要建立根节点, DOM4J 区别于 DOM 主要是因为对应接口不同, 其采用了 SAXREADER, 因此虽然时间复杂度相同, 但是实际开销比 DOM 小很多, 而 JDOM 是 DOM 方法在 JAVA 语言中的 API, 因此时间开销也比 DOM4J 大, 由此可得出 4 种解析方法的理论时间开销由大到小排序为 $C_{DOM} > C_{JDOM} > C_{DOM4J} > C_{SAX}$ 。

3 评价指标

以解析时间与 XML 文本大小的比值作为效率, 目前公认的评价指标是以效率的高低对实验结果进行判定。文献[21]利用解析时间为研究重点, 在相同大小的文件下对 DOM、SAX 解析方法进行比较, 得出 SAX 解析方法比 DOM 解析方法效率高的结论, 但存在 SAX 解析方法在价值取向上过于单一的缺点。

由于解析时间和 XML 文本大成正比例关系, 对结果的评价区分度不大。针对该问题, 本文以解析时间 $t(\text{ms})$ 、内存堆占用空间 $d(\text{MB})$ 、CPU 占用率 c 作为评价指标, 将时间和空间 2 个维度划分为分母和分子表示, 并分别进行加权, 加权和记为 T 、 M 。时间加权值指影响值 I 在时间维度上的加权值, 记作 β , 表示为分子的加权值; 空间加权值指对影响值 I 在空间维度上的加权值, 记作 α , 表示为分母的加权值。为了数据的归一化, 将 α 和 β 的取值范围设定在 0~1 之间, 并且控制 α 和 β 成反比例关系, 使结果受正比例关系的影响减小。影响值 I 的计算方法如下:

$$I(\text{影响值}) = (c \times 100 + d) \times \alpha / (t \times \beta) \quad (1)$$

$$\beta(\text{时间加权值}) = 1 - \alpha(\text{空间加权值}) \quad (2)$$

在式(1)中, 将影响值 I 作为解析方法优劣的判定依据, 主要分为以下 3 种情况:

1) 当时间加权值 β 和空间加权值 α 比重一样时, 影响值 I 作为直接影响值。

2) 以占用空间为重点, 即期望缩短解析时间(期望时间短则越优), 空间加权值 $\alpha > 0.5$, 在同等情况下, 若时间加权和 T 不变, $\lim_{M \rightarrow \infty} I = +\infty$, 则影响值 I 的值越大。

3) 以解析时间为重点, 即期望空间占用大(期望时间长则越优), 时间加权值 $\beta > 0.5$, 在同等情况

下, 若空间加权和 M 不变, $\lim_{T \rightarrow \infty} I = 0$, 则影响值 I 的值越小。

当不考虑时间和空间为价值取向时, 影响值 I 越小, 则解析方法越差; 当以缩短解析时间为重点时, 对堆内存容量和 CPU 占用率的要求高, 同时对处理时间的缩减要求也高, 因此此时影响值 I 值越小, 则该解析方法在此条件下越差; 相反, 以占用空间为重点时, 则对堆内存容量和 CPU 占用率要求低, 对处理时间缩减的要求也降低, 此时影响值 I 越大, 则该解析方法越差。以空间占用为重点(缩短解析时间为重点)时, 影响值 I 越大, 解析方法越好; 以解析时间为重点(占用堆内存空间为重点)时, 影响值 I 越小, 解析方法越好。在空间和时间加权和确定情况下, 加权值分配对影响值的影响如图 1 所示。

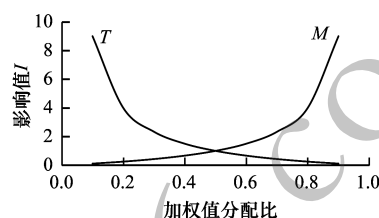


图 1 加权值分配比对影响值 I 的影响

Fig. 1 Effect of weighted value distribution ratio on effect value I

采用上述评价方法具有以下优点:

1) 将各个方法采用量值的方式进行比对, 尤其在数据量不大以及属性个数不多的情况下, 能够将解析方法对异构文本数据的转换结果影响程度划分, 评价指标以时间和空间划分的方式更加合理, 且计算简单。

2) 随着数据量增大, 不同解析方法的影响值差 I 值增大, 区分更加明显。

4 实验与结果分析

4.1 数据集

本文实验数据采用八爪鱼 V8.0 爬虫软件抓取电子商务网站用户 ID、名称、登陆地等基本用户标识及行为的日志信息, 并利用 .txt 文件对信息进行存储, 在 Eclipse 环境下将 .txt 信息转换成 XML 文件, 从而获得 Web 日志中 XML 文件资源, 该资源主要记录用户访问浏览网站的信息, 且具有数据量大、属性简单明确的特点, 数据总量为 2 207 620 条, 数据属性总共有 7 个。采用梯度划分法将数据分为 6 个梯度数据量的数据集, 并根据属性个数不同增加 4 个不同属性个数的数据集, 总共 10 份数据并按编号 1~10 进行排列, 数据集信息数量(条)的变化范围为 14 620~731 000, 属性个数(个)变化范围为

5~7,命名方式采用“Test_数据量_属性个数”,具体如表5所示。其中,编号8用于实验1、实验4和实验5,编号1、4、5、6、7、8用于实验2,编号1~编号3和编号8~编号10用于实验3。

表5 数据集的具体参数

Table 5 Specific parameters of data sets

编号	数据集名称	文件大小/MB	数据条数	属性个数
1	Test_1.4w_6	3.51	14 620	6
2	Test_1.4w_7	3.90	14 620	7
3	Test_1.4w_5	3.00	14 620	5
4	Test_14w_6	35.00	146 200	6
5	Test_28w_6	70.00	292 400	6
6	Test_42w_6	108.00	438 600	6
7	Test_56w_6	144.00	584 800	6
8	Test_70w_6	180.00	731 000	6
9	Test_70w_7	202.00	731 000	7
10	Test_70w_5	153.00	731 000	5

根据 DTD^[22] 规则,XML 文本的具体参数如表6所示。

表6 XML 文本的具体参数

Table 6 Specific parameters of XML text

属性变量	长度取值	描述
remote_addr	1	用户地址 IP
remote_user	4	用户名称
time_local	6	用户登录地
request	9	请求
status	10	请求判定
body_bytes_sent	7	浏览内容主体
http_referer	12	用户主机概况

4.2 实验环境

系统环境为 64 位 Win 10 操作系统,8GB 内存, Intel(R) Core(TM) i7 @ 2.40GHz。程序语言为 Java8.0, XML。实验工具为 Eclipse Java 4.9.0, Navicat10.7。数据库为 Mysql5.6.41。

4.3 实验过程

首先,进行“参数确定”的实验,其次,分别进行数据量、属性个数发生变化情况下的对比实验,以此确定其影响大小,最后,从占用空间和缩短时间角度分别进行实验,完成 4 种解析方法的对比。为了减少随机数据集带来的误差,所有实验均重复进行 10 次,并将所得的平均值作为最终结果。

4.3.1 参数确定

为了选取合理的时间和空间加权值,将 10 份异构数据分别转换为 XML 文本的半结构化数据,每份数据均采用 4 种解析方法对其解析并计算影响值 I 。实验通过区间分配权值的方法,取各种解析方法在 6 种不同数据量情况下影响值 I 的算数平均值,然后

分别在 5 个不同区间下进行计算,对比在不同的 α 和 β 组合情况下,影响值 I 的变化范围,实验结果如图 2~图 4 所示。由图 2 可知,以时间占用为重点, $0.8 \leq \beta \leq 0.9$ 时,影响值 I 的变化情况最为明显。由图 3 可知,以空间占用为重点, $0.8 \leq \alpha \leq 0.9$ 时,影响值 I 的变化情况最为明显。由图 4 可知,以数据量较大的 Test_70w_6 数据集为例,随着权值比例的降低,时间与空间之间的影响值 I 差值越来越小。根据文件大小及解析时间的差值分析,权重过大会导致结果不稳定。

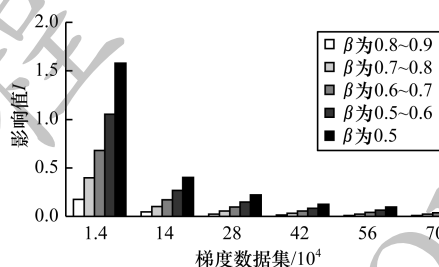


图2 以时间占用为重点时的影响值 I

Fig. 2 Effect value I with focusing on time occupation

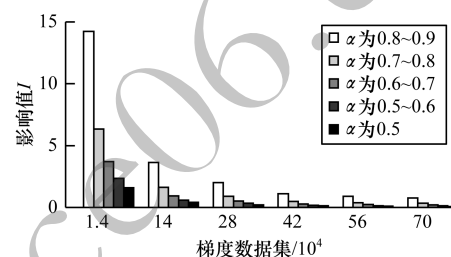


图3 以空间占用为重点时的影响值 I

Fig. 3 Effect value I with focusing on space occupation

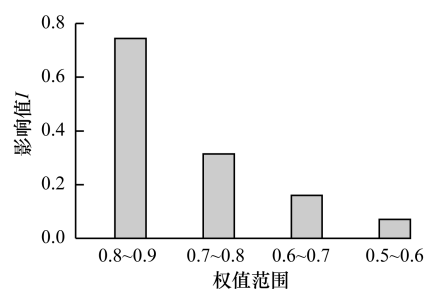


图4 在数据集 Test_70w_6 上影响值 I 的变化

Fig. 4 Change of the effect value I on the data set Test_70w_6

经过多次比对,以空间为重点时,本文选取 $\alpha = 0.85$, $\beta = 0.15$ 进行后续实验,以时间为重点时,本文选取 $\alpha = 0.15$, $\beta = 0.85$ 进行后续实验。

4.3.2 数据量对解析方法的影响

当数据量 n 从 1.4 万条逐渐变化至 70 万条时,在不考虑时间以及空间关系的情况下,实验比较了 4 种解析方法的影响值 I ,结果如图 5、图 6 所示,图 6 为图 5 中影响值 I 的局部放大。

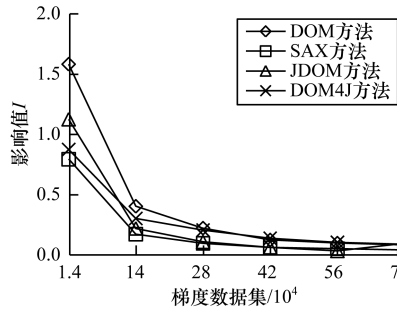


图5 4种解析方法在数据集 Test_1.4w_6 ~ Test_70w_6 下的影响值 I

Fig.5 Effect value I of four parsing methods under data sets Test_1.4w_6 ~ Test_70w_6

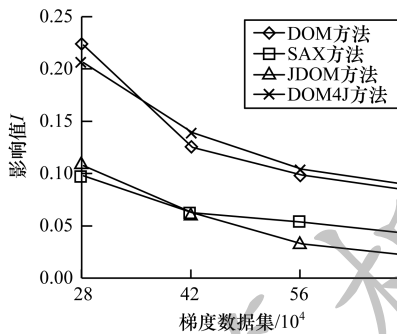


图6 4种解析方法在数据集 Test_28w_6 ~ Test_70w_6 下的影响值 I

Fig.6 Effect value I of four parsing methods under data sets Test_28w_6 ~ Test_70w_6

从图5和图6可得出以下结论:

1) 随着数据量 n 的增大,4种解析方法所花费的时间均增加,并且数据量 n 越大,4种解析方法的差距越小,因此本文主要针对数据量 n 增大的情况进行后续实验。

2) 当数据量 n 低于35万条时,DOM解析方法的影响值 I 比DOM4J解析方法高,数据量继续增大至高于35万条时,DOM4J解析方法的影响值 I 逐渐高于DOM解析方法,且始终比DOM解析方法的影响值 I 高。在时间和空间加权值比重相同时,DOM4J解析方法在35万条~70万条数据量时最优。

3) 在数据集 Test_42w_6 与 Test_56w_6 之间,SAX解析方法的影响值 I 逐渐高于JDOM解析方法,且与DOM4J、DOM解析方法的差值逐渐缩小,当时间和空间加权值比重相同时,SAX解析方法可能会随着数据量 n 的增大,影响值 I 逐渐升高。

实验结果表明,时间和空间加权值分配一致即不考虑时间空间影响时,当数据量 n 较小时,DOM解析方法具有很高的效率,继续增大数据量 n 时,该解析方法的效率反而降低,这是因为DOM解析方法采用的是树节点遍历全文的方式对文档解析,当数据量 n 较大时,建立树节点的时间会大幅增加,会造成效率变差,影响值 I 降低。

当数据量 n 较小时,从影响值 I 大小的角度分析,SAX解析方法的时间占用相对空间占用比其他3种解析方法多,且效果最差,当数据量 n 增大时,效率逐渐升高,且幅度较大。这是因为SAX解析方法采用的是流式处理文件的方式,逐行解析可以随时停止,不耗费空间资源,因此当数据量 n 增大时,影响值 I 会相对其他解析方法升高。

JDOM解析方法的影响值 I 随着数据量 n 的增大呈降低趋势,这是因为当数据量 n 增大时,在空间资源相差不大的情况下,JDOM解析方法耗费的时间远大于其他3种解析方法。

DOM4J解析方法的影响值 I 随着数据量 n 的增大呈升高趋势,这是因为虽然DOM4J解析方法采用获取根节点的方式遍历其子节点和属性,但是处理过程中可以根据接口选择SAX读取器,因此相对其他3种解析方法,其处理方式更快。

4.3.3 属性个数对解析方法的影响

实验对算法的时间复杂度进行分析,为了比较属性个数对解析结果的影响程度,在不考虑空间的影响因素下,以分子不变的影响值 I 作为判定依据。本文实验利用 Test_1.4w_6 和 Test_70w_6 数据集对属性加减,得到 Test_1.4w_7、Test_1.4w_5、Test_70w_7 和 Test_70w_5 这4个不同属性个数的数据集,比较6个数据集的影响值 I ,实验结果如图7和图8所示。

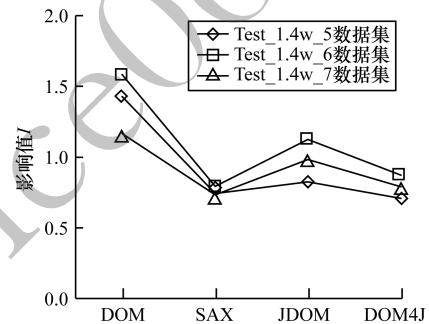


图7 4种解析方法在数据集 Test_1.4w_6 中的影响值 I
Fig.7 Effect value I of four parsing methods in the data set Test_1.4w_6

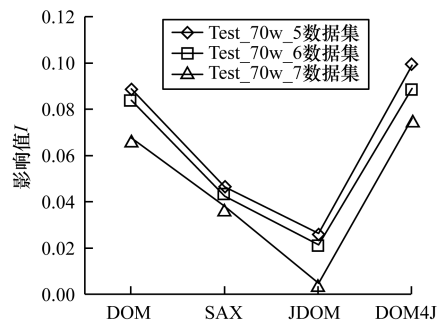


图8 4种解析方法在数据集 Test_70w_6 中的影响值 I
Fig.8 Effect value I of four parsing methods in the data set Test_70w_6

从图7和图8可以得出以下结论:

1) 在数据集 Test_1.4w_6 属性加减后的3个数据集上,当属性个数减少时,DOM解析方法的影响值 I 比

原数据集小;当属性个数增大或者减少时,JDOM 和 DOM4J 解析方法的影响值都比原数据集小。

2)在 6 个数据集上,属性个数的增加或者减少对 SAX 解析方法的影响值 I 影响不大,且影响值差值均低于其他 3 种解析方法。

3)在 Test_70w_6 属性加减后的 3 个数据集上,4 种解析方法的影响值 I 变化情况基本一致,当属性个数减少时,4 种解析方法的影响值 I 最高,而当属性个数增加时,4 种解析方法的影响值 I 最低。

实验结果表明,当数据量 n 不大时,属性个数的变化对结果的影响是非稳定性因素;当数据量 n 增大时,属性的变化对结果的影响是稳定性因素。因为在数据量 n 较小时,无论是解析空间还是解析时间,受实验环境的影响,限制条件被忽略,干扰性增强,当数据量 n 增大后,实验环境的影响对处理事件所耗费的时间干扰性减弱。由此表明,在不考虑时间和空间为重点并且数据量 n 较大的条件下,当属性个数增加时,影响值 I 降低,当属性个数减少时,影响值 I 升高。

4.3.4 缩短时间为重点时不同解析方法对比

本文实验将以 $\alpha = 0.15, \beta = 0.85$ 分别赋予空间和时间加权值。当考虑缩短时间为重点时,在数据集 Test_70w_6 上,对比 4 种解析方法影响值 I 的变化情况,并表示出该权重比与权重比为 1 时的差值,此差值表示为函数以时间为重点价值取向时的突出程度,结果如图 9 和图 10 所示。

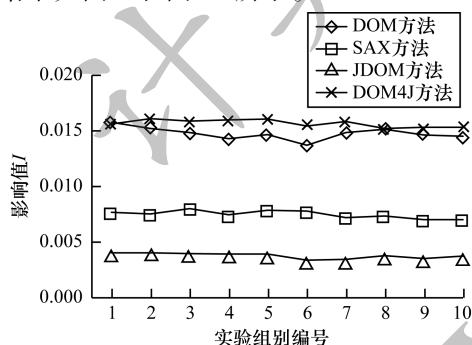


图 9 在数据集 Test_70w_6 中以时间为重点的 4 种解析方法的比较

Fig. 9 Comparison of four parsing methods with time-focused in the dataset Test_70w_6

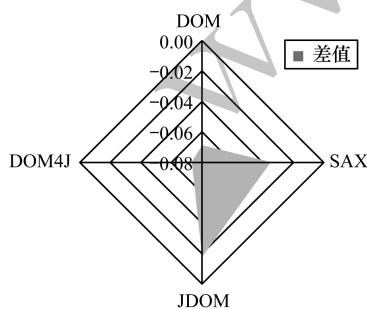


图 10 4 种解析方法的权重比差值比较

Fig. 10 Comparison of the difference between the weight ratios of the four parsing methods

从图 9 和图 10 可以得出以下结论:

1)DOM 和 DOM4J 解析方法在该条件下的影响值 I 均比 JDOM 和 SAX 解析方法高,且在 10 组实验中,DOM4J 解析方法的平均值大于 DOM 解析方法,说明以时间为重点时,DOM4J 解析方法的解析效果最好。

2)在差值对比分析过程中,JDOM 解析方法的差值最高(-0.017 67),说明在缩短数据解析时间上,JDOM 解析方法的时间开销最大。

实验结果表明,DOM4J 解析方法适用于以缩短解析时间为重点的实验环境。在实际的 XML 文本解析中,在实验条件较好且数据量 n 较大的情况下,DOM4J 解析方法的性价比最高,且实用性很强。

4.3.5 占用空间为重点时不同解析方法对比

本实验将以 $\alpha = 0.85, \beta = 0.15$ 分别赋予空间和时间加权值。在考虑占用空间重点时,同样在数据集 Test_70w_6 上,对比 4 种解析方法影响值 I 的变化情况,并表示出该权重比与权重比为 1 时的差值,此差值为函数以空间为重点时的突出程度。由于 DOM4J 和 JDOM 解析方法均是根据 DOM 解析方法中的数据存储方式来构建根节点树,需要把文档存至内存中,因此以空间为重点价值取向时,主要比较了 DOM 树结构和 SAX 流式处理文件结构,实验结果如图 11 所示。

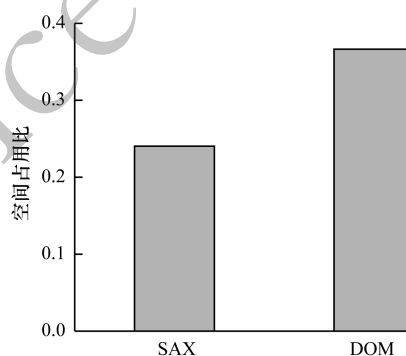


图 11 以空间为重点时的 SAX 和 DOM 解析方法的比较

Fig. 11 Comparison of SAX and DOM parsing methods with space-focused

从图 11 可以得出以下结论:

1)以占用空间为重点时,以 DOM 解析方法的树形结构方式空间占用比大于用 SAX 解析方法的流式处理文件结构,且在数据量 n 为 70 万条时,平均堆内存空间占用差值为 126 MB。

2)以树结构的存储数据的方式堆空间占用是流式文件处理方式的堆空间占用的 1.52 倍。

实验结果表明,当数据量 n 增大时,若考虑以空间占用为重点的实验环境,则采用 SAX 解析方法效果最好。

5 结束语

针对异构文本数据转换中,DOM、SAX、JDOM、DOM4J解析方法在不同情况下选择哪种方法更加有效,还存在缺乏科学实验论证的问题,本文提出对异构文本数据转换中XML解析方法进行对比研究,以3种不同评价指标来判定4种解析方法的优劣,得出了以解析时间为重点价值取向时,采用DOM4J解析方法最优,以空间占用为重点价值取向时,采用SAX解析方法最优的结论。但是本文也存在不足之处,如在进行不同梯度数据量的实验时,未针对数据量极大的情况进行实验,同时在CPU利用和空间占用的硬件利用问题上,也未尝试负载运行实验。在下一步研究中,拟选取数据量在100万条~500万条的XML数据集,调整数据梯度的纵向深度,对比观察数据量梯度变化与影响值的比例关系,推算解析算法的极限算力。而对于实验环境的硬件选择,拟采用集群式机群,采用分布式原理调整实验过程中的最大吞吐量,尝试负载运行机器,对比观察数据的处理时长,以期能够利用数据结果直观说明解析方法的内部处理方式。

参考文献

- [1] CHENG Hongtao. Research and implementation of unstructured text of data transformation into structured of data based on XML[J]. Modern Computer, 2013(9): 51-54. (in Chinese)
程洪涛. 基于XML的非结构化文本数据转换研究与实现[J]. 现代计算机, 2013(9): 51-54.
- [2] TARASOWA D, LANGE C, SORENN A. Measuring the quality of relational-to-RDF mappings[C]//Proceedings of International Conference on Knowledge Engineering and the Semantic Web. Berlin, Germany: Springer, 2015: 210-224.
- [3] WANG Lei, YAO Baofeng, ZHU Honghao, et al. Mapping method from XML to relational database non-affected by constraints of DTD changes[J]. Journal of University of Science and Technology Liaoning, 2011, 34(6): 588-593. (in Chinese)
王磊,姚保峰,朱洪浩,等. 一种无DTD变化约束的XML与关系数据库映射方法[J]. 辽宁科技大学学报, 2011, 34(6): 588-593.
- [4] LIU Jian, ZHANG Xiaoxiao. Dynamic labeling scheme for XML updates[J]. Knowledge-Based Systems, 2016, 106: 135-149.
- [5] COMBI C, MASINI A, OLIBONI B, et al. A hybrid logic for XML reference constraints[J]. Data & Knowledge Engineering, 2018, 115: 94-115.
- [6] REN Tingyan, LUO Gang. Functional dependency and inference rules for XML based on pivot nodes[J]. Computer and Digital Engineering, 2012, 40(1): 51-53. (in Chinese)
任廷艳,罗刚. 基于轴结点的XML函数依赖及推理规则[J]. 计算机与数字工程, 2012, 40(1): 51-53.
- [7] QIN Ying, MA Yongqi, MENG Lirong. Design and implementation of data conversion method based on XML[J]. Microcomputer & Its Applications, 2017, 36(20): 30-33, 38. (in Chinese)
秦英,马永起,蒙立荣. 一种基于XML的数据转换方法的设计与实现[J]. 微型机与应用, 2017, 36(20): 30-33, 38.
- [8] MARTENS W, NEVEN F, SCHWENTICK T, et al. Expressiveness and complexity of XML schema[J]. ACM Transactions on Database Systems, 2006, 31(3): 770-813.
- [9] QIU Xiaohua, QIN Shuanshuan, QIU Guo. Research on XML and JSON data transmission format based on WEB development[J]. Information Technology & Informatization, 2017(4): 123-125. (in Chinese)
仇小花,秦栓栓,邱果. 基于WEB开发中的XML与JSON数据传输格式研究[J]. 信息技术与信息化, 2017(4): 123-125.
- [10] LIU Zhiyong, ZHOU Jie, ZHANG Lin, et al. Research on TSP problem based on crossover and mutation combination[J]. Computer and Modernization, 2018(3): 54-59. (in Chinese)
刘志勇,周杰,张琳,等. 基于交叉与变异组合的TSP问题研究[J]. 计算机与现代化, 2018(3): 54-59.
- [11] CHEN Xiaomao, TANG Wenbing. Comparative research on methods of parsing XML in Java[J]. China New Technologies and New Products, 2009(15): 25. (in Chinese)
陈小毛,汤文兵. Java解析XML的方法比较研究[J]. 中国新技术新产品, 2009(15): 25.
- [12] ZHANG Jie. Research and application of parsing XML file in Java[J]. Silicon Valley, 2014, 7(6): 120, 128. (in Chinese)
张洁. Java解析XML文件的研究与应用[J]. 硅谷, 2014, 7(6): 120, 128.
- [13] LI Jian, MA Yanzhou. A webpage extraction method supporting visual configuration of DOM template[J]. Modern Computer, 2018(10): 56-60. (in Chinese)
李健,马延周. 支持DOM模板可视化配置的网页抽取方法[J]. 现代计算机, 2018(10): 56-60.
- [14] QIAN Chengyu, TANG Jianguo. Lightweight XML document parsing based on DOM tree[J]. Computer Programming Skills & Maintenance, 2016(18): 35-36. (in Chinese)
钱承聿,唐建国. 基于DOM树实现轻量级XML文档解析[J]. 电脑编程技巧与维护, 2016(18): 35-36.
- [15] LIU Yuxiao. Analysis and research of XML data analytical technique based on SAX[J]. Modern Electronics Technique, 2010, 33(12): 55-56, 65. (in Chinese)
刘雨潇. 基于SAX的XML数据解析技术分析研究[J]. 现代电子技术, 2010, 33(12): 55-56, 65.
- [16] HE Zhixue, LIAO Husheng. Query processing method of XML streaming data using list[J]. Journal of Computer Applications, 2016, 36(3): 665-669, 686. (in Chinese)
何志学,廖湖声. 基于列表的可扩展标记语言流数据查询处理方法[J]. 计算机应用, 2016, 36(3): 665-669, 686.

(上接第 293 页)

- [17] TIAN Yuan. Research on XML document analysis based on JDOM[J]. Computer CD Software and Applications, 2012, 15(6):112. (in Chinese)
田原. 基于 JDOM 的 XML 文档解析研究[J]. 计算机光盘软件与应用, 2012, 15(6):112.
- [18] WU Yan, TAN Xianhai. Research on heterogeneous data transformation based on XML [J]. Railway Computer Application, 2012, 21(10):4-7. (in Chinese)
武艳, 谭献海. 基于 XML 的异构数据转换的研究[J]. 铁路计算机应用, 2012, 21(10):4-7.
- [19] ZHANG Yifeng. Research on the analysis of DOM4j technology [J]. Modern Computer, 2011 (15): 39-42. (in Chinese)
张屹峰. DOM4j 解析技术探究 [J]. 现代计算机, 2011(15):39-42.
- [20] CHEN Feifei. Research and application of XML document parsing technology based on DOM4J [J]. Software Guide, 2016, 15(3):36-37. (in Chinese)
陈飞飞. 基于 DOM4J 的 XML 文档解析技术研究与应用[J]. 软件导刊, 2016, 15(3):36-37.
- [21] YANG Jing, ZHOU Shuang'e. Method for unstructured data transformation based on XML technology [J]. Computer Science, 2017, 44(z2):414-417. (in Chinese)
杨晶, 周双娥. 一种基于 XML 的非结构化数据转换方法[J]. 计算机科学, 2017, 44(z2):414-417.
- [22] JIAN Baorui, SONG Yuqing, CHEN Jianmei, et al. Model mapping scheme for XML documents based on RDBMS[J]. Application Research of Computers, 2011, 28(12):4621-4624. (in Chinese)
鉴保瑞, 宋余庆, 陈健美, 等. 一种基于关系的 XML 文档模型映射方法[J]. 计算机应用研究, 2011, 28(12):4621-4624.

编辑 刘继娟