5F-8

組込機器向け DOM のメモリ管理方式

羽藤淳平　佐々木幹郎　齋藤正史
三菱電機株式会社　情報技術総合研究所

1　はじめに

近年、インターネット上のコンテンツに限らず様々な分野で XML[2]が利用されており、組込み機器への適用も進んでいる。XML を処理する際の内部データ表現である解析木の構成要素として、DOM[2]を利用する場合が多いが、必要なメモリ量がコンテンツサイズに関連して増加する課題がある。組込み機器は PC と比較してリソース制約が大きく、DOM を含め組込み機器向けのソフトウェア開発でメモリ効率向上は大きな課題の一つである。上記の課題を踏まえ、我々は組込機器向け XML Parser を開発している[1]。

本論文では現在の XML Parser 内の DOM に関連するメモリの管理方式の問題点を挙げ、その問題点を解消するためのメモリ管理方式を述べる。

2　DOM メモリ管理方式 - 個別確保と一括確保 -

図 1 はある XHTML コンテンツを解析する場合に必要となるメモリサイズのグラフである。棒グラフ左から理論上最低限必要なサイズ、必要サイズを一回で必要最低限確保した場合のサイズ、全ての DOM インスタンスを個別にメモリ確保した場合のサイズである。一括確保と比較し、個別確保は 6.8 倍のメモリサイズを消費している。逆に、一括確保した場合は理論値に近い消費量となっている事が分かる。

以上の結果より、XML Parser では DOM インスタンス用のメモリ領域は、解析開始前に一定のメモリ領域を一括確保し、途中で不足した場合は一定量のメモリ領域を追加確保する方法を取っている[1]。

しかし、現在のメモリ一括確保の方法は、コンテンツの内容とは無関係に、確保するメモリサイズを決めている。そのため、図 1 の一括確保の様に理論値に近い消費メモリサイズとなる事は稀であり、メモリサイズを大きく設定すると、小さいコンテンツの場合に不必要なメモリ確保が発生し、逆に小さく設定すると、大きいコンテンツで何度もメモリ確保を行う課題がある。

Memory Management Method for Embedded DOM
Jumpei Hato, Mikio Sasaki, Masashi Saito
Information Technology R&D Center,
Mitsubishi Electric Corporation

3　提案方式

3.1　メモリ管理方式の課題

本章では 2 章で挙げた課題を解決するためのメモリ確保方式を説明する。この課題は解析結果である DOM 木によって消費されるメモリサイズが未知である事に起因する。DOM 木が解析した XML コンテンツの構造を反映する事から、消費メモリサイズを知るためには XML コンテンツ内の構成要素を数えれば正確に求める事が可能である。

しかし、XML 解析処理を二度実行する事になり、処理速度の面から実用的ではないため、コンテンツサイズを知る、もしくは推定する別の方法が必要となる。

その方法として我々は、過去解析したコンテンツのメモリ使用量を記録したメモリ使用ログを用いた方式と、その欠点を補う XML コンテンツの前解析の結果を利用する方式を提案する。

3.2　メモリ使用ログを用いた方式

図 2 はメモリ使用ログを用いたメモリサイズ決定方式のフロー図である。本方式は XML コンテンツ解析後、使用したメモリサイズを URL 等の XML コンテンツの識別情報、更新日時等と共にログに記録する。再度同一コンテンツを解析する場合にはログの情報から必要なメモリサイズを取得する事により、必要最低限のメモリサイズを取得する。また、ログに情報があったとしても、コンテンツ自体が更新される場合があるため、解析終了後に実際に使用したメモリサイズとログを比較し、異なる場合にはログを更新する。

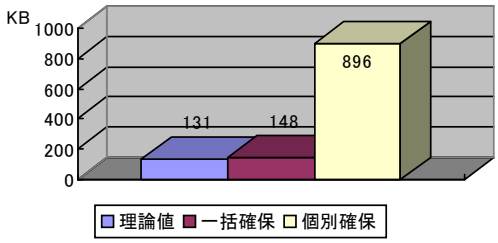

図 1 メモリ確保方式毎のメモリ消費量比較

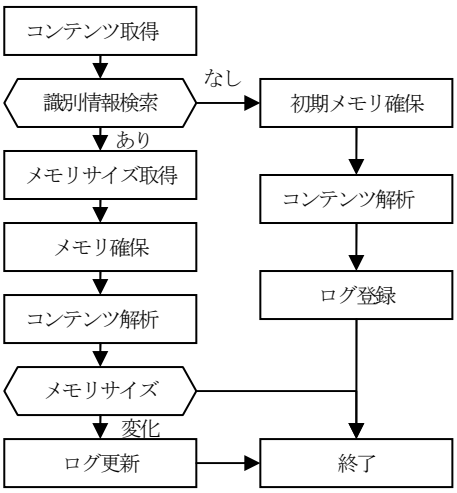

図 2 メモリ使用ログによるメモリサイズ決定法

本方式では、解析経験のないコンテンツを解析する場合にはメモリサイズを推定する事が出来ない欠点があるが、二回目以降の解析であれば、正確なメモリサイズを知る事が出来る。

3.3 コンテンツ前解析を利用する方式

本方式は、初回の解析時にのみ実行され、前解析を行い、コンテンツ構造を大まかに把握し、その情報を元に必要となるメモリサイズを推定､確保する。

本方式は前解析を行うため、従来方式よりも処理速度が遅くなる、正確なメモリサイズが求まるとは限らないと言った欠点がある。しかし、解析二回目以降はログによる方式のみで対応可能なため、常時発生する問題ではない。

3.4 前解析のパラメータ検討

前解析パラメータとして、コンテンツサイズや DOM 要素の個数などが考えられ、今回パラメータとしての妥当性を検討した。

表 1 は検討したパラメータと消費メモリサイズ間の線形近似式と、その R^2 値の表である。データは internet に存在する HTML コンテンツを使用し、サンプル数は 400 である。R^2 値は 0 から 1 の値を取り、1 に近い程よい精度で近似できた事を意味する。通常 0.8 以上あれば統計的によい精度だと判断される。よって、文字数、Comment 数以外のパラメータの R^2 値は 0.9 以上であり使用メモリサイズと十分な相関関係がある事が分かる。

次に図 3 は Element 数、Attr 数、Text 数、Comment 数、文字数にそれぞれ最小単位でのメモリサイズおよびそれぞれの R^2 値の積を重み付けした値と消費メモリサイズのグラフである。この場合の R^2 値は 0.9934 であり、単一パラメータの場合よりも高精度の近似であると言える。

以上の結果から、複数パラメータを用いる方法が一番よい近似であるが、コンテンツサイズ、単一の要素数でも統計的に十分よい近似が可能であり、実装環境の処理能力や必要とする精度等に応じてパラメータを使い分ける事が可能である。

表 1 各パラメータでの R^2 値

パラメータ	近似式	R^2
コンテンツサイズ	y = 0.0019x + 1.1789	0.9613
Element 数	y = 0.1033x - 0.3582	0.9237
Attr 数	y = 0.0598x + 2.8773	0.9502
Text 数	y = 0.0538x + 1.313	0.9785
Comment 数	y = 0.0036x + 4.5128	0.7995
文字数	y = 0.8377x + 16.696	0.3676

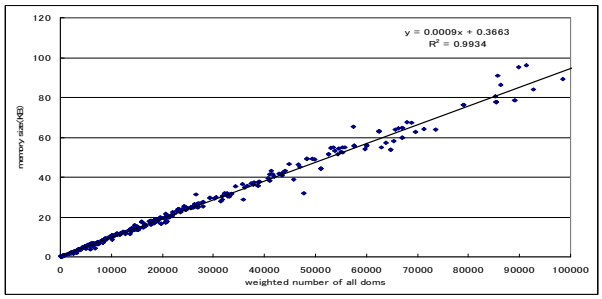

図 3 重み付け個数とメモリサイズ

4 おわりに

本論文で述べた、メモリ使用ログによるメモリサイズ決定方式とコンテンツ前解析を利用するメモリサイズ決定方式を組み合わせる事によって、従来では出来なかった、コンテンツ毎の内部構造等に合わせたメモリ確保が可能となる。

しかし、コンテンツ前解析で検討したパラメータは平均的に見てよいパラメータであると言えるが、図３から分かる通り、パラメータの値が大きくなるほど実際の消費メモリサイズのばらつきは大きくなっており、常によいパラメータとは言えない。そのため､今後はパラメータの更なる検討を行うと共に、本方式によるメモリサイズ決定法性能評価を行う予定である。

参考文献

[1] “組込み機器向け DOM モジュールの開発”, 羽藤 佐々木 齋藤, FIT2005
[2] DOM level1 specification
http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/
[3] XML specification
http://www.w3.org/TR/xml11/