オープンかつポータブルなデータベースガーベジコレクション
―データベースの不要なデータを同定・削除する―

1.背景

近年，メモリデータ構造をデータベースに保存するべくORMと呼ばれるオブジェクトデータベースマッピングツールが普及し始め、DB 内部のリンク構造は複雑になる一方である。

データ構造の管理に目を向けた場合、共有循環構造の削除は到達可能性予測が困難であるため課題となっている。 さらに安易な DB 上のデータ削除は DB 内部の制約違反を引き起こしかねないため、通常は削除を示すフラグを用いてデータを消さずに取り置く。しかしながらデータを消さずに取り置く方法ではデータが小さくことはないため，長期運用を前提とした情報システムでは不必要に高性能な DB を導入せざるを得なかった。また，データ流出等のリスクを開発ベンダーが取らなければならず，セキュリティー上にも問題を生じていた。

2.目的

ORMを使用したDBには決して参照されることがないことが保証されているデータが存在する。ORMはメモリ上のオブジェクトをそのままDBに格納するため、メモリ上ではGCによって削除されるようなオブジェクトもDBに保存される。そしてそのようなオブジェクトはデータベースの容量を無駄に占拠するだけの存在である。

本提案ではメモリ上で広く実現されているGCをDB上で実現することを目的とする。また特定の DB ソフトに依存せず動作するソフトウェアとする。

3.開発の内容

Java で標準的なORM であるHibernate 上でポータブルなデータベースガベージコレクション（DBGC）を実現した。

Hibernate をミドルウェアとして用いること DB のアクセスを抽象化できるので、特定の DB に依存せず GC を行うことが可能となる。

本プロジェクトの成果物はDBGCとDBGC 用に改変したHibernate である。DBGCはApacheライセンスを提供してオープンソースで公開した。ソフトウェアの誤った使用は DB を破壊しかねないため、教科書風のマニュアルをあわせて用意した。

GC のアルゴリズムにはマークアンドスイープ法を用いた。

ディスクへのアクセスはメモリアクセスに比べて高コストであるので、なるべく

減らしたい。それゆえに DBGC ではマーキングは SQL による集合演算によって行う。そのために必要なのはオブジェクトの到達可能経路情報とマーキングのための SQL の生成である。

Hibernate をはじめとした ORM ではその動作の原理上、DB のスキーマからエンティティ間の参照関係を格納した参照グラフを生成し、保持する。この参照グラフを解析することで追跡可能経路情報が得られる。

DB とのアクセスを最小限にするために、マーキングが収束するために必要なマーキング SQL の発行を最小回数にするような順序決定を行うためのグラフ解析アルゴリズムを作成した。

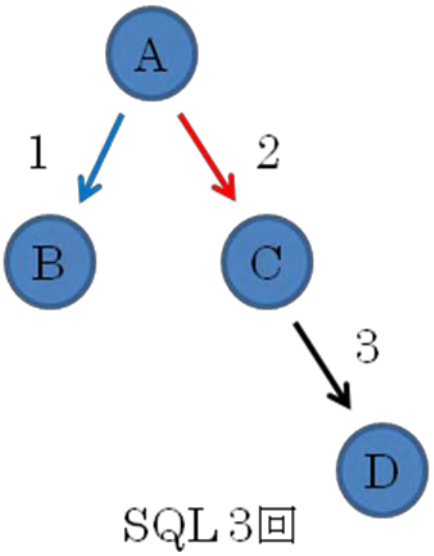

図1：循環参照がない場合のマーキング SQL の最適な発行順序。図中の円で表された A、B、C、D はそれぞれ DB 中のテーブルを意味し、矢印がマーキング SQL を意味する。

図1のようにオブジェクトに循環参照がない場合は、グラフで表されたテーブル（先述の追跡可能経路と等価）をトポロジカルソートすることで、マーキング SQL の最適な発行順序が決定できる。

循環参照がある場合は、グラフ中の循環参照のうち、他の循環参照がその部分集合とならないような循環参照をひとかたまりとみなすことで、図1のアルゴリズムを適用できる。ただしこのとき、循環参照までマーキングが行われたら、循環参照が収束するまでマーキングを行う。

DB を他のサービスと並列に行うために湯淺の snapshot-at-the-beginning アルゴリズムを用いた。本ソフトウェアでは GC のマーキングフラグとして GC を

生き残った回数を意味する整数値の世代番号を採用したこと。そのため、マークフェーズをマークアンドスイープの順序とは異なり、任意のタイミングで行うことができる。

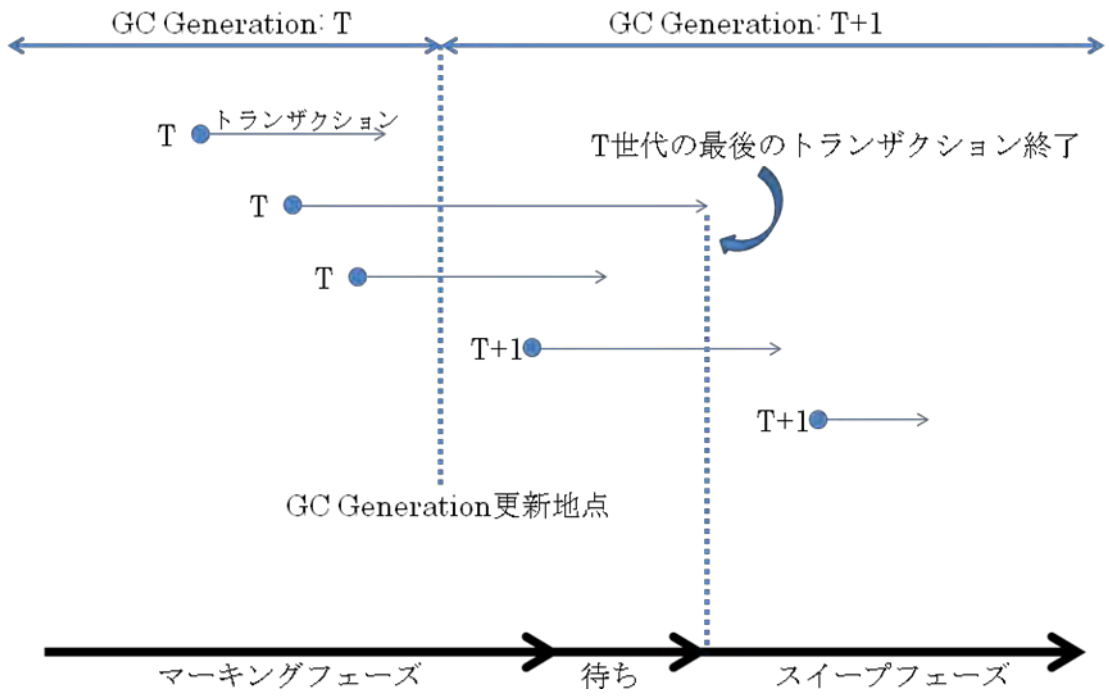

図2：DB で湯浅式 GC を実現するために必要なトランザクションの状態と、マークフェーズとスイープフェーズを実行する時期との関係を表した。

湯浅式 GC の実現のためにはユーザープログラムによる DB へのオブジェクトの新規追加とポインタの付け替えを実行時に監視し、その情報を得る必要がある。これは Hibernate が DB とのアクセスコストの削減のために持つメモリ上のオブジェクト監視機構に独自のコードを追加することで、実現した。

4．従来の技術（または機能）との相違

オブジェクトデータベースでの GC はいくつかの論文がある。また特定のオブジェクトデータベースでの GC ソフトウェアは存在する。

本ソフトウェアは、現在最も普及しているリレーショナルデータベースで GC が行える点と動作がデータベースソフトに依存しない点に特徴がある。

5.期待される効果

過剰な設備投資を躊躇う中小規模の EC サイトへの導入が期待される。

また個人情報を扱うシステムを運営している企業に対しては、情報流出などのセキュリティリスクの回避のみならず、契約で明記されている個人情報の保存期間後のデータ削除コストの削減に本ソフトウェアの効果が期待される。

6.普及（または活用）の見通し

現在は、本ソフトウェアの機能の一部を Hibernate 本体に取り込むための活動を行っている。個人的なつながりがあるシステムで運用実績を稼ぐための準備をしている。

7.開発者名（所属）

郷原浩之　東京大学大学院工学系研究科システム創成学専攻