2P-5

電子掲示板バックアップシステムの開発

峯岸　純也　　早川　栄一
拓殖大学　工学部　情報工学科
negi@os.cs.takushoku-u.ac.jp　　hayakawa@cs.takushoku-u.ac.jp

１．はじめに

携帯電話からアクセスできるなどの機能の充実や、手軽に自分用のインターネットのページを持つことができることもあり、掲示板や blog のレンタルサービスの利用者が増えている。

レンタル掲示板の問題点は次の 3 点である：(1)サーバの故障や停止が行われるとデータの復元ができない、(2)サーバのデータベースやファイルに直接アクセスできないものがほとんどであり、記事データの保存が行えない、(3)掲示板、blog のレンタルサービスでは記事のバックアップ機能がほとんど存在しない。

そこで、本報告では掲示板や blog の記事部分を自動的に抽出し、XML 形式でバックアップ、再利用することを目的としたシステムの開発について述べる。

２．特徴

(1) 記事データ抽出

HTMLデータからXML文書への完全自動変換は現在ではほとんど不可能[1]である。そこで本システムではユーザが書込み設定を行い投稿することで、掲示板データから記事データを抽出する。そして可搬性向上のために記事データ（投稿者、題名、本文など）をXML形式に変換し、保存を行う。

記事データの取得方法としては、はじめにバックアップする掲示板の記事部分のタグ構造をダウンロードした掲示板データと、書込み設定を行った掲示板データとの差分抽出より取得する。取得したタグ構造によって、記事データの位置が判断が行える。

次に掲示板の記事データの取得には、差分抽出により得られたタグ構造とバックアップを行う掲示板データとの比較から、書き込まれた記事の部分を取得する。

(2) 自動バックアップ

バックアップを行う操作をユーザが行っていると、バックアップすることを忘れてしまった時に、書き込んだデータが消えてしまい、復元することができなくなってしまう。そこで自動的なバックアップの方法として、本システムでは、一定期間で自動的にバックアップを行う方法を実装するほか、次に示す方法も実装する。

RSS ファイルを利用した自動バックアップとして、一定期間でバックアップを行う掲示板の RSS ファイルを取得する。取得を行った際に RSS ファイルの更新が行われていたら、自動的にバックアップ処理を行う。また取得を行った際の RSS ファイルを保存し、更新状況の確認を本システムから行えるようにする。

RSSPing を利用した自動バックアップとして、blog が更新された時に更新された情報が本システムに送られてくる。送られてくる情報を利用して、自動的にバックアップを行うと共に、更新情報を本システムで作成し、ユーザに配信することが可能になる。

３．記事データの抽出方法

本システムではユーザが書込み設定を行うことで XML 文書へ変換して保存する。その際に行う差分抽出手法を図 1 に示す。

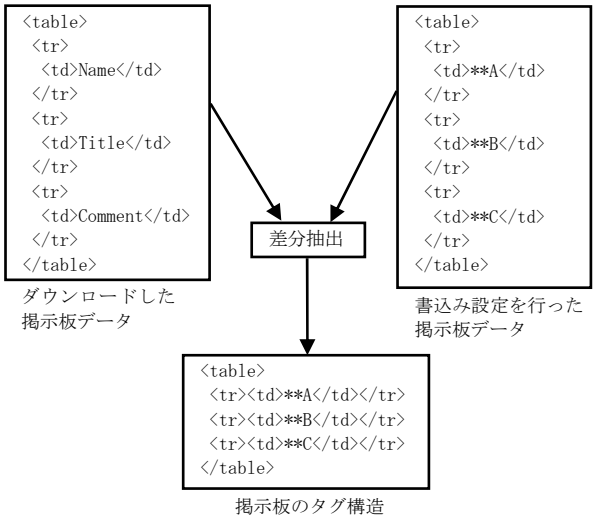

図 1　差分抽出手法

(1) 書込み設定

書込み設定とは、取得する記事データの項目ごとに本システムであらかじめ決めた文字列（書込み設定文字列）を入力し、投稿することである。図 1 では**A、**B、**C がそれぞれ投稿者、題名、本文の書込み設定文字列に対応している。書込み設定を行うのは、バックアップする掲示板のデータをダウンロードした後になる。

ユーザによって書込み設定文字列の投稿を行ったデータが、図 1 の書込み設定を行った掲示板データとなる。よって書込み設定を行った掲示板データには、ダウンロードした掲示板データも含まれている。

(2) 差分抽出

ユーザが書込み設定を行った後、本システムが行う処理として、図 1 の差分抽出処理を行う。

差分抽出処理は、ダウンロードしたデータと書込み設定を行ったデータの差分を取得する処理を行う。図 1 で

Development of a Electronic Bulletin Board Backup System
Junya Minegishi, Eiichi Hayakawa
School of Engineering, Takushoku University

差分として取得できる部分は、新たにユーザが書込み設定のために投稿を行った部分となる。差分抽出によって得られたタグ構造がバックアップする掲示板の記事データ部分のタグ構造となる。

(3) 記事データ取得

差分によって取得したタグ構造中の書込み設定文字列の位置によって、掲示板データと比較を行う際に、記事データが存在しているのか判別が可能になる。このことから、差分抽出より取得した掲示板のタグ構造とダウンロードした掲示板データと比較を行うことで、記事データの取得を行い、XML のタグ付けを行って XML 形式で保存を行う。

4. 設計

全体構成を次の図 2 に示す。

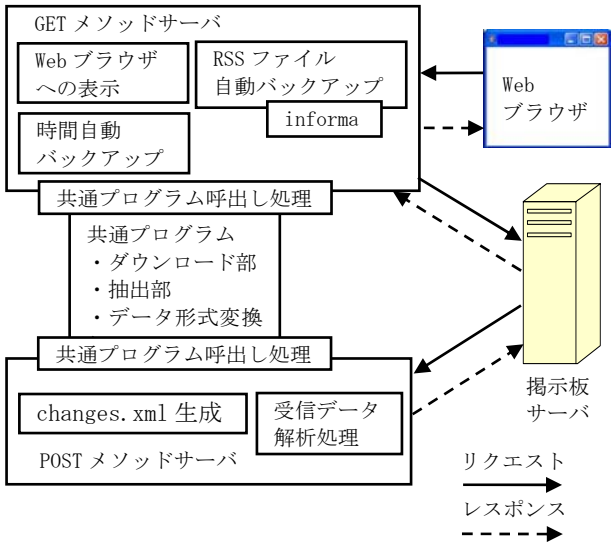

図 2　全体構成

(1) GET メソッドサーバ

GET メソッドサーバの具体的な処理としては、Web ブラウザからのアクセスの要求によって、時間での自動バックアップやバックアップ処理を行うために共通プログラム呼出し処理を行うものである。

RSSファイルでの自動バックアップでは、掲示板サーバに定期的にリクエストを送り、RSSファイルを取得し、informa[2]というRSSファイルの解析・生成を行うJavaライブラリを使用してバックアップ処理を行う。

(2) POST メソッドサーバ

POST メソッドサーバの具体的な処理としては、送られてきたデータが RSSPing の形式であれば解析、共通プログラムの呼出し、バックアップ処理を行う。

また送られてきた RSSPing 形式の blog の更新情報から changes.xml として保存し直し、本システムで各 blog の更新情報を配信する。

(3) 共通プログラム

共通プログラムは各サーバプログラムから呼出し処理があった場合に、バックアップ処理を行うプログラムである。具体的な処理内容は、ダウロード部ではバックアップする掲示板のデータを本システムに保存する、抽出部ではダウンロード部で保存したデータと、差分により得たタグ構造を比較し記事データの抽出を行う、データ形式変換部では抽出された記事データにタグ付けを行う。これにより XML 文書に変換して保存をする。

5. 実現

本システムは Java 言語で開発を行い、約 4500 行のプログラムになっている。

本システムの評価として 5 社の掲示板を借りて、実際に記事データの抽出を行った。

5 社うち 3 社については、差分抽出によって、掲示板の記事データ部分のタグ構造が得られたので、記事の取得が可能であった。

残りの 2 社のうち 1 社は、あらかじめ掲示板のタグ構造を用意すれば、記事データの取得は可能であった。

最後の 1 社については、最初の 1 件の記事データは取得が可能であったが、掲示板の仕様で 2 件目以降の記事のタグ構造が、差分抽出より得たタグ構造と大きく変わるので、2 件目以降の記事の取得が困難であった。

本システムの実行画面を次の図 3 に、抽出した記事データの XML 文書を次の図 4 に示す。

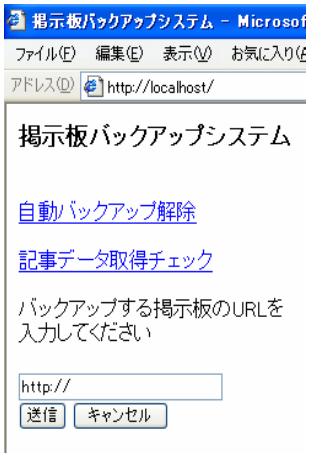

図 3　実行画面

図 4　XML 文書

6. おわりに

本報告では、掲示板や blog に関するバックアップシステムの実現について述べた。抽出した記事部分の HTML データの XML 文書化および、一定時間ごとの RSS ファイル、RSSPing での自動バックアップ機能の実装を行った。

今後の課題は、差分抽出の改良、取得できる記事データ項目の追加、記事データのデータベース化である。

参考文献

[1] 岩沼宏治　事例に基づく HTML 文書の XML 文書への自動変換システムの構築（継続）
http://www.taf.or.jp/publication/kjosei_18/pdf/073.pdf

[2] informa (RSS API)
http://210.224.170.171/java/other/rss_informa.html

[3] 松本和之　構造化アノテーションを用いた知識再利用性の高い電子掲示板
http://www.nagao.nuie.nagoyau.ac.jp/papers/matsumoto_ipsj67.xml