

## WWW 情報の保守管理用ウェブエンジンの構築

5 R-5

王樵 朱仲武 近藤邦雄

埼玉大学大学院理工学研究科

## 1 はじめに

WWW (the World Wild Web) は急速の発展によって、今は全世界での大きな情報空間となった。従って、WWW 情報 (ウェブページなど) を直接自分の情報として処理できるアプリケーション [1,2,3] も増えてきた。これらの情報は分散的に世界中に存在し、ダイナミックに更新されているので、その保守管理は大変困難である。このような WWW 情報を保守管理できるソフトウェアがあれば、大変便利である。

本研究では、ウェブロボットの基本アルゴリズムを利用し、アプリケーションで利用される特定の WWW 情報の保守管理 (データの自動登録と定期更新、関連情報の修正、保守情報の提供) の立場から、WWW 情報の保守管理用ウェブエンジンを提案する。そして、できるだけ多くのアプリケーション開発者に使用してもらうため、汎用性を考えながら、そのウェブエンジンをパッケージ化して構築する。最後に、構築されたウェブエンジンを学習支援システム WWWCALIST で応用し、その実用性を評価した結果について述べる。

## 2 ウェブロボットについて

ウェブロボットは、単純な基本アルゴリズム [4] を使って、自律的に WWW 上のリソースを探索収集するソフトウェアであり、主に WWW 検索サービスの中で使われている。ウェブロボットを用いる WWW 検索サービスは、ウェブロボットの使い方によって、さらに off-line 型と on-line 型の二つのタイプに分けられる。

off-line 型とは、サービスを開始する前に予めウェブロボットを使ってウェブ情報を収集するタイプの WWW 検索サービスである。off-line 型の場合、常に最新の情報をユーザに提供することや増え続ける巨大化するデータベースを維持することは困難である。off-line 型の欠点を解決するため、情報収集はユーザの要求があるたびに実行する on-line 型 WWW 検索サービスが開発されたが、検索時間が長く、ウェブの一部しか検索対象にできず、実用例が少ない。

また、ウェブロボットを用いる検索サービスは、Yahoo! のようなイエローページ型 WWW 検索サービス

と比べると、検索の精度が遥かに悪く、ウェブロボットがアクセスされたウェブサーバに負荷をかけて、崩壊させる危険性もある。従って現在は、ウェブロボットを専門化するアプローチや、ウェブロボットのアクセス先を制限する動きなどがみられる。

## 3 ウェブエンジンの構築

## 3.1 ウェブエンジンの位置付け

ウェブロボットが自動的にウェブページを収集できる性質を活かして、アプリケーションで利用される WWW 情報の保守管理での活用が考えられる。また、これらの WWW 情報が特定されていて、また使用目的 (WWW 検索でなく特定 WWW 情報の保守管理) も違うので、WWW 検索サービスで発生した問題が避けられる。既存のウェブロボットと区別するため、このようなソフトウェアは WWW 情報の保守管理用ウェブエンジンと呼ぶ。本研究では、特定のアプリケーションだけに対応するのではなく、パッケージとして汎用的に利用するウェブエンジンを提案する。つまり、図 1 に示すように、アプリケーション開発者は提供されたウェブエンジン (パッケージ) を自分用に拡張 (中間データの変換、関連情報の修正、保守情報の処理) して、アプリケーションに組み込む。アプリケーションが実行されると、ウェブエンジンが WWW からそのアプリケーションのための情報を抽出し、中間データとして提供する。その中間データがアプリケーション用に変換された後、データベースに登録する。ウェブエンジンが定時的にそのデータベースを更新するが、その際、もし変更があれば、アプリケーションの中の他の関連情報を自動修正する。そしてソフトウェアだけが解決できない保守問題もあるので、データ収集の成否、データの変更などを保守情報としてシステム管理者に提供する。

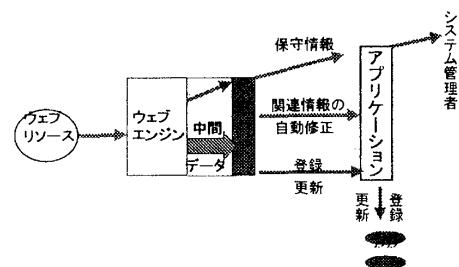


図 1: ウェブエンジンとアプリケーションとの関連

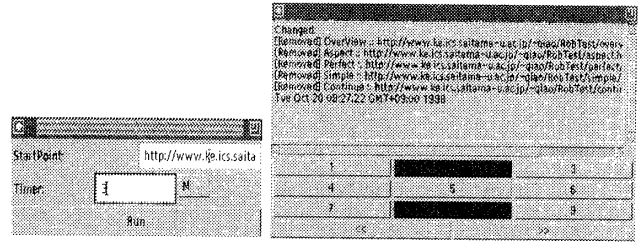
<sup>1</sup> Construction of WebEngine for Maintaining and Managing WWW Information

<sup>2</sup> Qiao Wang, Zhongwu Zhu and Kunio Kondo

<sup>3</sup> Graduate School of Science and Engineering, Saitama University

### 3.2 ウェブエンジンの機能

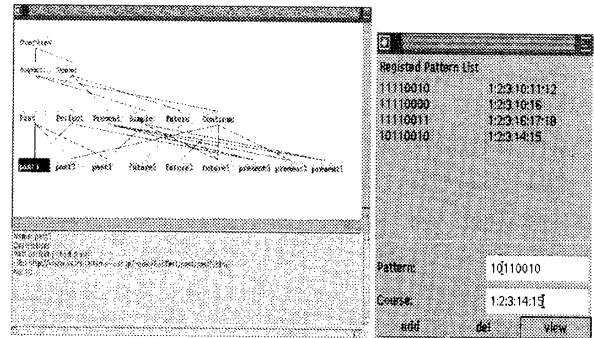
- (1) 定時保守：アプリケーションから与えられた時間間隔で保守管理を行なう。
- (2) ウェブ渋滞防止策：一つのウェブページに対し、複数回アクセス（10秒ごとに5回まで）をする。
- (3) パッケージとしての機能：  
アプリケーション開発者が簡単にウェブエンジンを自分用に拡張して、アプリケーションに組み込むため、共通データとして提供される中間データをアプリケーション用のデータへの変換処理を定義する中間データ変換メソッド、データの登録更新や関連情報の自動修正の処理を定義するデータ登録更新と関連情報の修正用メソッド、データ収集の成否やデータの変更などの保守情報に対する処理を定義する保守情報処理メソッドが提供される。



(a) ウェブエンジン設定 (b) 保守情報

### 3.3 ウェブエンジンの特徴

ウェブページからどんな情報を抽出するかによって、異なるウェブロボットが開発されたが、WWW情報の保守管理用ウェブエンジンを構築する際も、どんな中間データをアプリケーション開発者に提供するかが最重要である。既存のウェブロボットで採用されるリスト型データを用いると、各WWW情報要素間の関連を必要とするアプリケーションに対応できなくなるので、本研究では、データの完全性と汎用性を考えて、スタートポイントからなるツリー型構造の中間データを採用した。各ノードには、そのウェブページのURL、タイトルと存在する画像のURLリストが含まれる。



(c) マップ1 (d) 知識登録1

図 2: システム実行

## 4 学習支援システムへの応用

### 4.1 概要

WWWCALIST[1,2,3]とは、本研究と同時に開発したWWWにおける個人適応型学習支援システムである。WWW CALISTでは、まずウェブエンジンが教材の全体情報を取得し、中間データをマップデータに変換してデータベースに登録する。つぎに、教師が教材の全体のマップを見ながら、1、0からなる学生の回答パターンを予測し、学習ガイド（学習すべきマップのノード番号列）を知識情報として登録する。ウェブエンジンが定期的に教材マップを更新し、変更があれば、関連の知識情報をも自動修正する。そして、定期的に保守情報を保守情報データベースに保存する。

### 4.2 結果

図2の(a)でウェブエンジンのスタートポイントとデータ更新の時間間隔を設定し、実行すると、(c)のような教材全体マップが作られ、これに基づいて教師が(d)で知識情報を登録する。データ更新があれば、その知識情報を自動修正したり、(b)のような保守情報をシステム管理者に提供する。学生がWWWCALIST[1,2,3]を使用するとき、オンラインテストした後、サーバから送り出される教材のマップに沿って学習していくことができる。

## 5 まとめ

本論文では、アプリケーションで利用されるWWW情報の保守管理の立場から、WWW情報の保守管理用ウェブエンジンをパッケージ化して構築した。そして、WWW-CALISTに応用し、その実用性を評価し、有効性を示した。本ウェブエンジンは、イメージデータ管理やウェブサイドの可視化などへの応用も可能である。

### 参考文献

- [1] 朱仲武、王樵、近藤邦雄、バルーズファー、CBRに基づく学習状態に応じた知識関連マップによる学習支援、1998年度人工知能学会全国大会（第12回）、PP.606-609
- [2] Zhongwu Zhu, Qiao Wang, Behrouz H. Far, Kunio Kondo: "A Web-based Individualized Adaptive Computer Aided Learning System", WebNet 98-World Conference of the WWW, Internet and Intranet (1998)
- [3] Zhongwu Zhu, Qiao Wang, Behrouz H. Far, Kunio Kondo: "WWW-CALIST: A General Purpose Tool for Constructing Web-based Individual Adaptive CAL Systems", WebNet 98-World Conference of the WWW, Internet and Intranet (1998)
- [4] 田村健人、WWWの検索サービスは何をしているのですか？情報処理38巻12号、1997年12月