4T-04 エージェント技術を利用したWeb情報取得・表示システムの提案

金子郁夫*,　松永賢次**,　綿貫理明**
*専修大学大学院 経営研究科 情報管理コース
**専修大学 経営学部

1.研究の目的

WWWで何か情報を得ようとする場合，必要な同種の情報がいくつかのページに分散していたり，関連している情報が別のページに存在したりすると探索および閲覧に時間がかかる。そこで，ユーザーが必要とするWWW情報をより短時間で閲覧するために，ユーザーの望む情報と関連のあるＷｅｂ情報を収集・加工して１つにまとめて表示するシステムを提案する。

2.システム全体のアーキテクチャ

このシステムでは，ユーザーが求める情報に関する知識をもとに検索し，取得したファイルのそれぞれに対してレイアウトを調べ，設定された基準でパーツに分解して抽出されたものを加工・表示する。

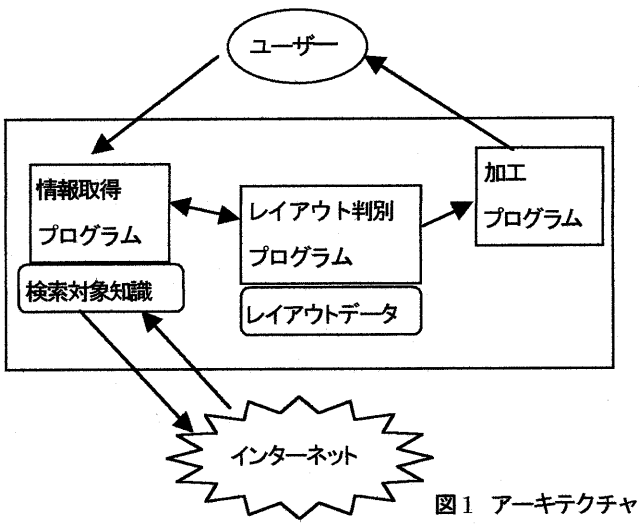

図1　アーキテクチャ

A Proposal of Web Information Acquisition and Display System Using Agent Technology
Ikuo KANEKO, Kenji MATSUNAGA and Osaaki WATANUKI
Business Administration and Commerce, Graduate Schools of Senshu University
Department of Information Management, Senshu University
2-1-1 Higasimita, Tama, Kawasaki, Kanagawa, 214-8580, Japan

2.1 情報取得プログラムと検索対象知識

ここでは，検索対象知識をもとに検索と情報取得を行う。レイアウト判別プログラムには選別された情報を渡すため，ここでページの選別を行う。選別の戦略は，最初は検索対象知識とマッチする量で決めて，以降はレイアウト判別プログラムで抜き出された必要な部分のレイアウトや内容との比較によっても選別を行う。

この部分が保持している情報は，「検索対象に関する知識」と「検索を行ったページとその中で有用であると判断されたページのＵＲＬ」である。

検索対象知識はユーザーがカスタマイズできるもので，検索したい内容についての情報を単語列で記述したものである。この知識を変えることによって検索対象や情報取得の戦略を変えることができる。これに関しては，ユーザーに何もない状態から作らせるだけでなく，ユーザーが選んだモデルページから半自動的に抜き出せる機能も持たせる。

取得するページのＵＲＬに関しては現在のところ検索対象や検索範囲を絞って研究を行うことで検索範囲の拡大や適切な情報源を抜き出すことの失敗を考慮しないことにしているので，事前に有用だと判断されたページのＵＲＬを持っている状態から始まることになる。

2.2 レイアウト判別と情報の取捨

このプログラムでは情報取得プログラムから渡されたファイルのレイアウトを調べ，情報の取捨選択を行なう。ファイルからの抽出処理は機械的にできなければユーザーとの対話を通じて必要な部分を選び出す。

レイアウトは，<TABLE>と<IMG>それぞれを１つのブロックとみなし，それ以外は段落や区切りを表すタグ（例えば
や<PRE>）を目安として１つのブロックとして扱う。検索対象知識と照ら

し合わせて単語がマッチした部分に関係すると考えられる名詞や数字は優先的に必要な部分とみなす。たとえば，テーブルの一番左側の列や最上段の行にキーとなる文字列が記述されている事が多い。この事を考慮していくつかのパターンを用意しておくことで，キーに対してどの文字列が関連していてどれを抜き出すかをユーザーに問う手間が省け，同一形式で書かれているページについても流用できる。

通常の文章は，たとえば，文章の途中でないと判断される＜Ｈ１＞タグなどで強調された１行の文字列とその後に続く文章を関連するひとまとまりとみなして，その最初の文字列をキーとして，抜き出す内容はその後に続く文章とする方法が考えられる。それ以外で判断が困難なもの（画像など）はユーザーに判定してもらう。

ユーザーに選択してもらう方法は，切り出したブロックを画面上にウィンドウで並べて必要なものを選択してもらう。更に細かい選択に対応するために，テーブル表現の場合は「テーブル全体」「横一列」「縦一列」「セル単位」での選択が可能なようにしたい。それ以外の文章は全体が必要なのか一部が必要なのかを考えると難しいので，かなり大雑把な選択ができるにとどまってしまう。

レイアウトデータは，元のファイルからどの部分を抜き出したかを示す情報を付加した元データと，処理しやすいように変換されたデータの双方を保持している。もとのデータを保持しておくのは，情報取得プログラムが情報源ファイルの更新をチェックするためである。したがってこのレイアウトデータは情報取得プログラムも参照する。

2.3 加工プログラム

加工プログラムではレイアウト判別プログラムで作成されたレイアウトデータをもとに適切な形に加工する。

加工したデータを表示する形式は，ウィンドウを左右に分割し，左側をツリーやリストなどのキーワードを一覧で視覚的に比較できる表現にし，右側に左側で選択したキーに結びつく内容を表示する形式を考えている。

例えばパソコンの価格情報を表示したいとすると，左側にパソコンの型番か画像をリストで表示し，左側にスペックや価格を表示するようにしたい。

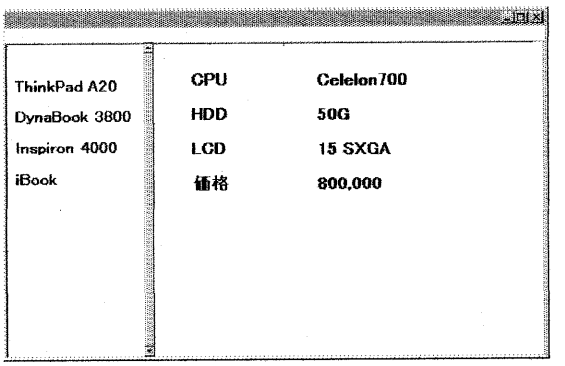

図2 パソコンのスペックと価格を例とした表示

キーとなる語を選別する方法は，検索対象知識に書かれている単語が候補になる。その中からキーとする語を選択する。事前にどの単語をキーとするかを決めておけるようにもする。この際，各キーに対してどのコンテンツが結びついているのかという関連を示す情報が必要となる。この情報はレイアウトデータをもとにつくり，加工プログラムで保持しておく必要がある。

3.考察

情報取得プログラムで現在のところは情報源をかなり限定しなければならないので汎用性をもたせるためにはどうしたらよいかを考える。

レイアウトの判別では，ＨＴＭＬ記述のファイルから必要な情報を抽出することがとても困難なので，ページの作者自身が，現行のＨＴＭＬから付加情報を記述できるＸＭＬ等へ移行することが望ましい。

加工プログラムにおいてキーとコンテンツの関係を表現する方法を考える必要がある。

参考文献

1) 松永賢次：インターネット上に分散した数値情報を取得するエージェントソフトウェア，専修経営論集，No. 69, pp.135-157, Dec.(1999).

2) Hermans,B. : Intelligent Software Agents on the Internet, http://www.hermans.org/agents (1996).