情 報 学 基 礎 42－3
（1996. 7. 25）

知識に基づくインターネットからの情報獲得と統合化

岩爪道昭　白神謙吾　畑谷和右
武田英明　西田豊明
奈良先端科学技術大学院大学 情報科学研究科
〒630-01 奈良県生駒市高山町 8916-5

我々は, 広域ネットワーク上での情報氾濫の問題に対処するため, 特定の対象領域に関する基本語彙の体系 (オントロジー) を利用して, インターネット上に散在する情報を自動的に収集・分類するシステム IICA (Intelligent Information Collector and Analyzer) の開発を進めてきた.

本研究では, IICA に, WWW ページからの情報抽出・統合化機能を追加するために, 2 種類の情報抽出法, (1) 状態遷移図による方法, (2) 特徴記述ルールによる方法を考案し, HEDIR システムとして実装した. 評価実験の結果, 状態遷移図による方法で正解率約 70 %, 特徴記述ルールによる方法で, 適合率が約 80 %となり, 我々のアプローチが, WWW の多様な分野の情報獲得・統合化に有効であることが明らかになった.

Knowledge-Based Information Capturing and Reorganization from the Internet

Michiaki IWAZUME, Kengo SHIRAKAMI, Kazuaki HATADANI, Hideaki TAKEDA, Toyoaki NISHIDA
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-01 Japan

The number and diversity of infomation is increasing rapidly. So, it is difficult to select necessary information. In this paper, we propose a system of extracting information using heuristics on the WWW.

We propose two methods. One is the method using automaton, the other is the method using rules of description. In the former, we divide a page to sentenses, first. Then, we extract a sentense that includes information and divide the sentense to words. Finally, we decide the word that is a startpoint and translate words with automaton. In the latter, we divide a page to sentenses and a sentense to words, first. Then, we extract information that match with a rule. Futhermore, we integrate the two methods. we translate sentenses with the former after extracting with the latter

As a result, we extract infomation 70 % exactly with the method using automation. we extracts infomation with 60 % recall and 80 % precision when we use the method using rules of description. The result of the experiments indicates that the approach of this paper provide us a new style of using WWW.

1 はじめに

近年, インターネット上で提供される情報は多様化・複雑化・大規模化の一途をたどっており, 個人で処理しなければならない情報の量は, 人間の処理能力の限界をすでに超えてしまった. 我々がネットワークから必要な情報や知識を獲得するためには, 収集・整理・理解の各過程において多大な時間と労力を費やさなければならない.

このような情報氾濫の問題に対応するため, ネットワーク上では, 様々な検索エンジン, ツールのサービスが行なわれている. 例えば, Alta Vista [1] では、 2100 万の WWW ページを収集していると言われており, このようなツールを利用すれば, 情報収集は比較的たやすく達成するように思われる.

しかし, 既存の検索ツールでは, 大量の検索結果が未整理のまま出力され, その中から所望の情報を探し出すことにも時間と労力が必要になる.

大量の情報の理解を支援するためには, 従来の情報の収集・検索のみではなく, オントロジーをはじめとする知識に基づく情報分類, 情報抽出, 情報統合 (再組織化) といった, 情報の内容に立入ったより知的なシステム必要である.

すでに, 情報の内容処理に関する研究 [2], [3], [4] が幾つかなされているが, WWW のような多様な分野の情報源に対応できるでけの汎用性を目指したものはない.

我々はこれまで, 特定の対象領域に関する基本語彙の体系 (オントロジー) を利用して, 広域ネットワークに散在する情報を自動的に (1) 収集, (2) 分類を行なう IICA (Intelligent Information Collector and Analyzer) と呼ぶシステムを作成してきた [5].

しかし, 現実には, これら 2 つの機能のみでは, 特定のカテゴリーに大量の WWW ページが収集・分類された場合, 内容の理解を十分に支援することは難しかった.

そこで, 我々は, オントロジーのクラスごとに定義した, キーワードやフレーズに着目した情報抽出ルールによって, WWW ページから該当する記述部分を自動抽出し, 収集した複数のページの情報を統合化して表示する (3) 内容抽出・統合化機能を IICA に追加・実装した. 図 1に, IICA によって収集・分類された温泉に関する WWW ページから温泉名, 最寄り駅, アクセス方法, 風呂の種類, 泉質の項目を抽出・統合化した結果例を示す.

本論文では, 収集・分類した WWW ページから、特定の情報を抽出し行なう IICA の情報抽出エンジン HEDIR (Heuristic Extraction from Distributed Information Resourses) について述べる. さらに, HEDIR の評価実験について紹介し, 我々のアプローチが WWW 上の多様な分野における情報獲得および統合化に有効であることを示す.

2 情報抽出システム HEDIR

2.1 HEDIR の概要

HEDIR の概要を図 2に示す. HEDIR は, IICA による分類処理済の WWW ページを入力とし, WWW ページに各分野関する情報がどのように記述されているかヒューリスティックスを利用して情報抽出を行ない, 抽出・解析結果を該当する情報として抽出する.

HEDIR の処理フローは, 情報記述の部分を抽出した後, そのまま抽出結果を出力するケースと抽出結果をさらに解析処理を行うケースに大別される. 前者は, 例えば、温泉の効能といった特定の単語が抽出できれば十分な場合に適用される.

一方, 後者は, 観光地への交通手段のように, 記述部分を単に抽出するだけでは不十分場合に適用され, 状態遷移図に

URL	温泉の名前	最寄り駅	アクセス方法	風呂の種類	泉質
asse-spa-j.html	"赤湯温泉"		"バス"		"炭酸水素"
hitagi-spa-j.html	"日奈久温泉"	"JR八代駅"	"JR日奈久駅下車"		"食塩泉" "単純"
kamiseto-spa-j.html	"金桜温泉"	"JR三角駅"	"バス"		"炭酸水素"
kurogiyama-spa-j.html	"鶴木山温泉"	"JR佐敷駅"			"単純"
kuroyu-spa-j.html	"鶴湯温泉"		"徒歩"		"単純"
yoshio-spa-j.html	"吉尾温泉"	"JR吉尾駅"	"徒歩"		"単純"
yuno-ko-spa-j.html	"足温泉"	"JR水俣駅"	"バス"	"分合いの湯"	"重曹泉"
yunotsuru-spa-j.html	"鶴温泉"	"JR水俣駅"	"バス"		"単純"
yunoura-spa-j.html	"湯浦温泉"	"JR湯浦駅"	"徒歩"		"単純"

図 1: 温泉情報の統合化例

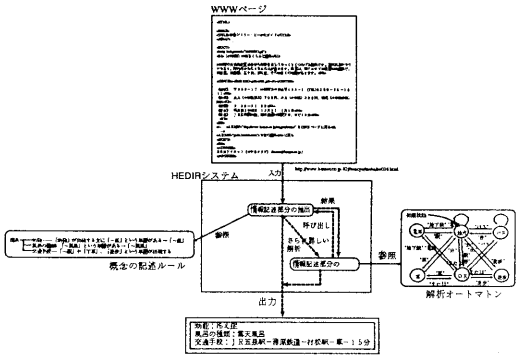

図 2: HEDIR の概要

表したルールを利用して記述部分の解析を行い, その結果を出力する.

3 情報抽出手法

内容を自動抽出する手法として, 1) 特徴記述ルールを利用した方法, 2) 状態遷移図を利用した方法, の 2 つを提案する. 以下の節でこの 2 つの方法についての詳細を述べる.

3.1 HTML の基本処理

HTML フォーマットで書かれた WWW ページに対する処理として, HEDIR では (1) 文単位への分割, (2) 単語単位への分割の 2 つの処理を行っている. ここで, 文単位への分割は, 記述の内容が変わる箇所で分けており, 分割された記述部分の 1 つ 1 つを文と呼ぶことにする.

記述内容の差異によって WWW ページを文単位に分割することで, 文の 1 つ 1 つがどのような内容かがはっきりする. 本研究では表 1を基準に WWW ページを文単位に分割した.

(2) では, JUMAN[6] を用いて形態素解析し, 単語の部分マッチングによって, 情報抽出を行っている. さら、形態素解析を行なった後, 表 2のルールを用いて単語を連結させることで, ユーザの意図に近い解析結果が得られるようにしている.

(1) 「<Hn> ~ </Hn>」のパターンで記述してあれば, それを1つの文と判断(n = 1 or 2 or 3)
(2) 「<DT> ~ </DL>」, 「<DT> ~ <DT>」, 「<DT> ~ <P>」, 「<DT> ~ <DL>」, 「<DT> ~ <UL>」のパターンで記述してあれば, それを1つの文と判断
(3) 「<LI> ~ </UL>」, 「<LI> ~ <LI>」, 「<LI> ~ <P>」, 「<LI> ~ <DL>」, 「<LI> ~ <UL>」のパターンで記述してあれば, それを1つの文と判断.
(4) その他のタグについては, 文とみなさず無視
(5) その他の場合は, 通常の文章のように「. 」を文の区切りとする.

表 1: 文単位への分割に関する基準

<1> 数字が隣接している場合
<2> 数字の直後に単位が出現する場合
<3> アルファベット同士が隣接している場合,
<4> 固有名詞 or 普通名詞が隣接している場合

表 2: 形態素解析の後処理に関するルール

3.2 概念の記述ルールを用いた方法

ここでは, オントロジーの各概念に対応した, 情報抽出ルールを用いた方法について説明する.

観光情報を提供するWWWページでは, 温泉情報を例に調べると, 「効能は神経痛である」, 「露天風呂がある」といった記述が頻繁に出現する. そこで, ユーザの目的や趣向に合った情報を提供するために, ルールをあらかじめ設定し, 簡単な言語表現パターンを利用した, 情報の抽出を行った.

記述を取り出すルールの設定手順 必要な情報の記述箇所を抽出ためのルール設定は, 次の手順で行なう.

1. 属性情報の設定
   例えば, ある温泉のページについて調べる場合, オントロジーの温泉の概念に関する属性情報として, 温泉名, 効能, 泉質, 風呂の種類などを設定する. 図3に定義記述例を示す. この例では, 「**温泉**は, 値を1つ取る**温泉の名前**という属性と, 値を1つ以上取る**風呂の種類, 泉質, 効能**という属性をもつ**訪問地**である」と定義されている. ここで, ただ1つの値を持つ場合には *has-one* , 複数の値を持つ場合には *has-some* という述語が用いられている. また, *is-a* (「~はである.」),

```
(define-pclass (温泉 ((has-one 温泉の名前)
                      (is-a 訪問地)
                      (has-some 風呂の種類)
                      (has-some 泉質)
                      (has-some 効能)
                      )))
```

図 3: 温泉に関する属性情報の定義

```
(define-concept (効能 (is 傷病 with (or "効能>" "効果" "効く
"))))
(define-concept (傷病 (or "+ 症>" "+ 傷>" "+ 病>")))
```

図 4: 温泉の効能抽出のためのルール

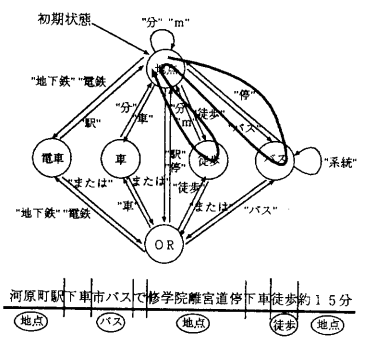

図 5: 状態遷移図を利用した交通手段記述の解釈

has-at-least(「少なくとも~を持つ.」) といった述語も使うことができる.

2. 各属性情報特有の言語表現パターンを記述
   実際にテキストから抽出を行なうルール部分の設定を行なう. 図4に, 温泉の効能に関する抽出ルールの例を示す. 最初のルールは, 「"効能" または "効く" という単語と同一文に, **傷病**という概念があれば, それば効能である」ということを意味し, 2つ目のルールは「"+ 症" または "+ 傷" または "+ 痛" という表現パターンが存在すれば, 該当する内容が**傷病**である」という意味である. ここで, "+" は1以上の長さの文字列を表す記号で, "+ 症" の場合は, "冷え症", "高血圧症" などの表現にマッチする.

現在, 温泉, 寺, 神社, 飲食店等の情報が掲載されているWWWページについて,

1. オントロジーの概念の選択
2. その概念の属性の選定
3. 各属性に対応するテキストの記述部分の解析

という手順で, 手作業で行なっている. ルールの設定自体は, どの概念でも図4のような形式で統一的に記述可能で, 属性の数も多い場合で10数個程度である.

3.3 状態遷移図を利用した方法

テキストの中で, 特定の単語の後にどのような単語が出現するかが経験的に予測できる場合に, 状態遷移図して文の解析を行なう方法ある.

3.3.1 本方法の手順

図5は交通手段記述の解釈を行うための状態遷移図である. この状態遷移図を利用して観光地の交通手段情報を抽出する例について説明する.

まずはWWWページを文単位に分割し, 文単位で情報抽出が行えるようにする. 次に, 求める情報が記述してある文を抽出する. その次に, その文に対して形態素解析を行う文を単語単位に分割にすることで, 状態遷移図を利用した記述

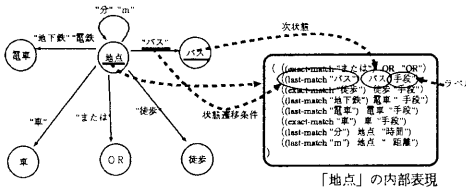

図 6: 状態遷移図の内部表現の例

解釈ができるようになる．ここまでは，状態遷移図を利用した方法ための前処理であり，これにより図5にある記述部分が抽出される．

前処理の次に，その文中のどの単語から状態遷移図による解釈処理を始めるかを決定する．図5の例では記述部分の最初にある駅名から始めるとしているので，「河原町駅」から始めることにする． その後に，状態遷移図を用いて記述文の意味解釈を行う (図5参照)． 状態遷移図を用いて記述文の解釈処理をしていく内に，最終状態に達したと判断されれば解釈処理を終了する．本論文での最終状態の判定基準は，(1)10単語連続して状態遷移が起こらない，(2) 文の終わりに達する，(3) 現在の状態から遷移する先がない，の３つである．解釈処理が終われば結果を出力する (図5では「河原町駅－市バス－修学院離宮道停－徒歩－約１５分」)．

次に，その状態遷移図の構造について説明する．

3.3.2 状態遷移図の構造

ここでは状態遷移の内部構造について説明する．１つ１つの状態の内部構造は基本的にリスト構造からなっている (図6参照)．真になる状態遷移条件がなければ，記述文の次の単語に対して状態遷移条件の判断を行う．

例えば「～バス」という単語パターンに一致した場合に条件遷移させたいときは，(last-match “バス”) と書く．本論文では，状態遷移図を書くための述語の定義については詳しく触れない．(詳細は [7] を参照)．

次に，「次状態」には次の状態の名前が書かれている．

最後に，「ラベル」は条件にマッチした単語がどのような種類であるかを示すためのものである． 例えば，研究室のWWW ページにおいて研究室メンバーを抽出した際に学生と教官を分けて出力する際にこのようなラベルを利用すればよい．

4 情報抽出手法の統合化

本章では、提案した２つの手法の統合化について述べる．

ここでいう，情報抽出手法の統合化とは、概念の記述ルールを利用した手法を用いて情報記述の部分を抽出し、状態遷移図を利用した手法を用いて記述部分の解析を行うことである．

4.1 統合化された処理の手順

ここで、概念の構成要素を、1) 記述ルールにマッチする文を抽出した場合に抽出結果をそのまま出力する、2) 記述ルールにマッチする文を抽出した場合に状態遷移図を利用した解釈処理を行う、の２つの種類に分ける．図7に統合化された処理の手順を示す．

最初に、Web ページ全体を文単位に分割し、さらに形態素解析を行う． 次に、各構成要素の記述ルールにマッチした文を抽出する．その次に、1) の構成要素の場合はそのまま結果を出力し、2) の場合は状態遷移図を用いて意味解釈を行い、その結果を出力する (図7参照)．

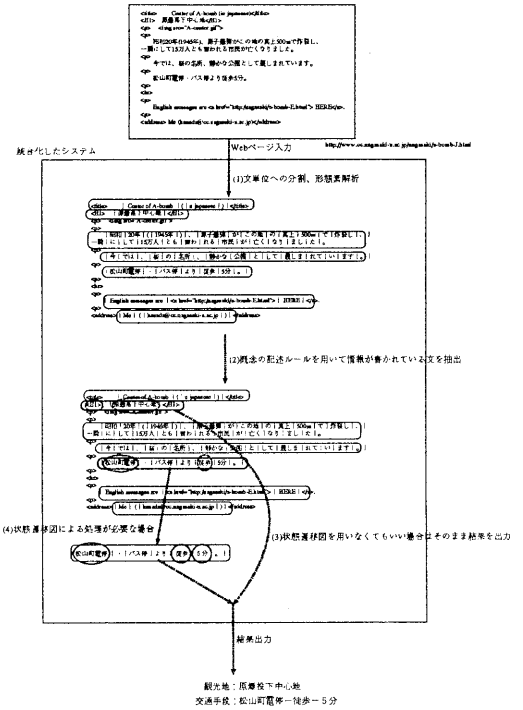

図 7: ２つの手法を統合化した場合の処理フロー

5 評価実験

5.1 特徴記述ルールによる方法の評価

あらかじめ分類したWWW ページを，温泉・飲食店・寺についてそれぞれ100 件用意し，1) それぞれの分野に対してルールを記述する，2) WWW ページ全体を文単位に分割し，形態素解析をする，3) WWW ページ全体にルールの適用を試み，ルールにマッチした記述だけを抽出する，の手順で実験を行った．

表 3に実験結果を示す．

表 3の意味するところは以下の通りである．評価の基準については，1 つ 1 つのページに対して正しく抽出された item 数と本来抽出するべき item 数を数え，再現率 (正しく抽出された item 数 / 本来抽出するべき item 数) と適合率 (正しく抽出された item 数 / 抽出された全 item 数) を計算した．それぞれの分野及び全体の平均を計算した．

実験の結果，簡単なヒューリスティックスを使っただけにも関わらず比較的高い精度で情報が抽出できている． しかし，これからどのようにすれば，さらに高い精度で情報抽出できるかを考える必要がある．

表 3の結果を 3 分野平均で見ると，再現率が 6 割，適合率が 8 割だといえるが，飲食店の再現率が極端に低い．その原因は，飲食店のページには料理名が多く見られるが，現在利用している知識では全ての料理名を網羅できなかったことである．また，構築したシステムでは知識獲得ができないので，知識が不足している場合に対処できないという短所がある．その短所を補うためのシステムを設計する必要がある．

分野	適合率	再現率
温泉	82.2 %	61.2 %
寺	72.2 %	73.4 %
飲食店	85.0 %	41.0 %
3 分野の平均	79.8 %	58.6 %

表 3: 実験結果

1. 正確に記述部分を抽出したページ	85/100
2. 正しく記述部分を解析したページ	70/100

表 4: 実験結果

その他に実験の結果から考えられる問題点は (1) 複数の寺, 温泉, 飲食店が書かれている WWW ページに対する対策が不十分, (2) 文の区切り方, 単語の区切り方がこちらの意図と違う, (3) ほんの少しルールと違うだけでも情報を抽出できない, の 3 つである.

(1) については, 本方法では 1 つのページに複数の温泉（もしくは寺, 飲食店）が書かれている場合は考慮しておらず, そのような構造に対応するために構造解析する必要がある.

(2) については, より正確な文単位への分割, 形態素解析を行うためのヒューリスティックスの確立が必要であろう.

(3) の解決策としては, ルールと記述がどの程度マッチしているかの適合度を設け, 適合度がある基準を越えていれば情報として抽出するというシステムを設計することが挙げられる.

5.2 状態遷移図を利用した方法

交通手段の書かれている WWW ページを研究者自身が 100 件収集し, それらに対して, 1) WWW ページを文単位に分割, 2) それぞれの WWW ページに対して, 交通手段記述のある文を抽出, 3) 抽出した文に対して形態素解析, 4) 前節までに述べた方法で作成した状態遷移図 5 を用いて文の解釈, の手順で実験を行った.

交通手段が記述されている文を抽出する時に判断するためのヒューリスティックスとして, (1) 「<Hn> 交通 </Hn>」(n = 1 or 2 or 3), 「<DT> 交通 <DD>」, 「交通 ：」という記述の直後にある, (2) 駅名らしき単語を含む, の 2 つを用いる. 後者については, 1) 「～駅」という単語, 2) 全国駅名データベースにある単語と完全一致する単語, という 2 条件のどちらかにマッチした単語を駅名と判断している.

状態遷移図については, 10 件のサンプルページを参照して状態遷移図を研究者自身が作成した. 以上の手順で実験を行った結果は表 4の通りである (実験の典型例は図 5).

1. は再現率を意味しており, それに対して適合率は 82.3 % (= 70/85×100) である. 正確に記述部分が抽出された割合については, 2 つのヒューリスティックスだけを利用したにもかかわらず, 85 % という比較的高い正解率が得られた. 一方, 残り 15 % の誤りの原因はヒューリスティックスの不足であることがわかった.

ヒューリスティックスの不足の典型例は図 8である. この WWW ページには, 交通手段が書かれていることを示すタイトルもなければ, 駅名らしき記述も見られない. このようなヒューリスティックスを用意しておけば解決する可能であるが, そのためには, 新しいヒューリスティックスの獲得を支援する機能が必要である.

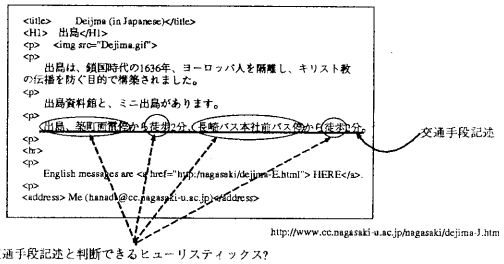

図 8: ヒューリスティックスが不足していた例

1. 正確に記述部分を抽出したページ	88/100
2. 正しく記述部分を解析したページ	75/100

表 5: 実験 1 の結果

次に, 正確に記述部分が解析されたページの割合は 70 % と比較的高い率が得られた. 一方, 30 % の誤りの内 15 % は 1. において正確に記述部分を抽出できなかったからである. 残りの 15 % の誤りの内訳は, 1) 状態遷移図による表現の限界が 9.5 %, 2) ヒューリスティックスの不足が 5.5 %, である.

1) については, 状態遷移図による表現は十分に柔軟であると考えて実験を行ったが, 実際に実験を実施すると, 同義語で記述されているなどの表現のできない部分が少なからず見られた. よって, 同義語の処理については考慮する必要があることがわかった.

2) については, 正確に記述部分を抽出できなかった原因と同じことが言え, ヒューリスティックスを獲得できるような枠組が必要だと思われる.

5.3 情報抽出手法の統合化の評価

ここでは 2 つの実験を行った. 実験 1 は 5.1 節との比較実験である. 実験 2 は観光地の Web ページばかりでなく、他の分野に対しても統合化させた手法は有効であるのかを確かめる実験である.

5.3.1 実験 1

ここでは、5.2 節と同一の 100 の Web ページに対して、1) テキストを文単位に分割、2) 概念の記述ルールを用いて交通手段の記述文を抽出、3) 抽出した記述部分に対して形態素解析、4) 状態遷移図を用いて記述部分の意味解釈をする、の手順で実験を行った. 状態遷移図の作成方法は 5.2 節と同一である.

実験の結果は表 5の通りである (例は図 7). 表 5は何を表しているについては、5.2 節の実験と同じで再現率を表している. 再現率は 75 % で、適合率は 85.2 % であった.

5.2 節の実験と比較して再現率が向上したということは、5.2 節で用いたヒューリスティックスを利用するよりも概念の記述ルールを利用した手法の方が正確に情報抽出が行えるということである. その理由としては、Web ページ全体に対して形態素解析を行っている点にあると考えられる.

分野	適合率	再現率
研究室	80.5 %	63.5 %
イベント	65.6 %	70.8 %
2 分野の平均	73.1 %	67.2 %

表 6: 実験 2 の結果

5.3.2 実験 2

ここでは情報系の研究室、イベントに関するページをそれぞれ 10 件づつ用意して、それらに対して実験を行った.

情報系の研究室、イベントそれぞれの概念に対してルールを設定し、10 の Web ページに対して統合化した手法を適用した. ルールはそれぞれの分野に対して半分の 5 ページを参考にして作成した. イベントについては日時、研究室については構成員に関して状態遷移図による処理を行っている.

評価の基準は 3.2.3 項と同一である. 分野ごとの平均と 2 分野での平均を表 6に記す (例は図 9).

実験の結果を見てみると、3.2.3 項の実験と差のない結果が得られており、本論文で提案した手法が観光地以外の分野にも適用可能であると思われる.

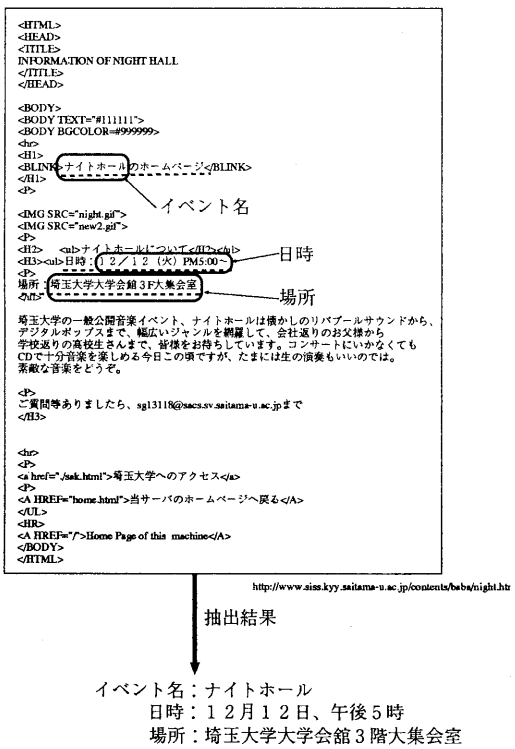

図 9: イベントのページの情報抽出例

5.4 考察

実験の結果, 我々が提案した方法には次のようなメリットがあることがわかった.

(1) 提案した方法は, 分野に関係なく安定した精度で情報抽出が可能である.

(2) 状態遷移図を利用した方法の大きな特徴は, 1 つの状態遷移図の適用範囲は狭いが, 深い情報抽出処理が行えることである.

(3) 基本語彙の体系 (オントロジー) を意識したスロット構造は情報抽出に有効利用できる.

(1) については, 分野によって少し差はあるものの, 合計 5 分野 (温泉, 飲食店・寺・研究室・イベント)の WWW ページからの情報抽出において, 7 割前後の再現率・適合率が得られた. 研究室・イベントの WWW ページについては 10 件づつしか実験を行っていないが, すべてのページにおいて安定した再現率・適合率が得られた. よって, 分野に関係なく安定した精度で情報抽出できると考えられる.

6 まとめ

本研究では, インターネットからの情報収集・分析システム IICA に, WWW ページからの情報抽出・統合化機能を追加するため, 2 種類の情報抽出法すなわち, (1) 状態遷移図による方法, (2) 特徴記述ルールによる方法を提案し, HEDIR システムとして実装した. 評価実験の結果我々のアプローチが, WWW の多様な分野の情報獲得と統合化に有効であることを示した. 今後は、情報抽出ヒューリスティックスの獲得支援に取りむ予定である.

参考文献

[1] http://www.altavista.digital.com

[2] 住田一男, 知野哲朗, 小野顕司. 文書構造解析に基づく自動抄録生成と検索提示機能としての評価. 電子情報通信学会論文誌 D-II, Vol. J78-D-II, No. 3, pp. 511–519, 1995.

[3] 佐藤円, 佐藤理史, 篠田陽一. 電子ニュースのダイジェスト自動生成. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2371–2379, 1995.

[4] 松尾利行, 武田英明, 西田豊明. 技術情報空間の構築と探訪の知的支援に関する研究. 信学技報 AI95-33, Vol. 95, No. 265, pp. 87–94, 1995.

[5] 岩爪道昭, 武田英明, 西田豊明. オントロジーに基づく広域ネットワークからの情報収集と分類. 情報研報, vol. 95, p25-32, 1995.

[6] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム JUMAN 使用説明書 version 2.0. 1993.

[7] 畑谷和右, ヒューリスティックスを利用したWWWからの情報抽出. 修士論文, 奈良先端科学技術大学院大学情報科学研究科, 1996.