テキストからの情報抽出・統合化法の提案と知的情報収集・分析システム IICA の実験的評価

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

電話: 07437-2-5265, fax: 07437-2-5269

e-mail: {mitiak-i, kengo-s, kazuak-h, takeda, nishida}@is.aist-nara.ac.jp

近年, インターネットに代表される情報ネットワーク上で提供される情報は急速に多様化, 大規模化しており, 人間の処理能力では対応が困難になってきている. 我々は, 広域ネットワークに散在する大量の情報群から, オントロジーと呼ばれる対系知識を用いて, 必要な情報を自動的に収集・分類するシステム IICA の研究開発を進めている. 本稿では, 収集した情報の内容理解をさらに支援するために, 領域特有の言語表現パターンに着目したヒューリスティックスによるテキストからの情報抽出・統合化法を提案する.

さらに, WWW を対象した IICA の情報収集・分類・抽出の各機能に関する実験的評価を示す. 評価の結果、我々のアプローチがネットワーク上の大量かつ多様な情報に適用可能な頑健性と実用性を備えていることがわかった.

キーワード: オントロジー, インターネット, WWW, 情報収集, 情報分類, 情報抽出・統合化

Heuristic Exstraction and Reorganization of Information from Texts, and Evaluation of an Intelligent Information Collecting and Anlyzing System: IICA

Michiaki IWAZUME, Kengo SHIRAKAMI, Kazuaki Hatadani

Hideaki TAKEDA and Toyoaki NISHIDA

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-01 Japan

phone: 07437-2-5265, fax: 07437-2-5269

e-mail: {mitiak-i, kengo-s, kazuak-h, takeda, nishida}@is.aist-nara.ac.jp

In this paper, we present a system called IICA (Intelligent Information Collector and Analyzer) which gathers, classifies, and reorganizes information from wide-area network. Ontology plays an important role in IICA. It specifies the common background knowledge shared by the user and IICA, allows IICA to make inexact match between the user's request and the candidates, and assigns user-oriented categories. IICA extracts information using a state transition network grammar and concept frames. We have implemented and evaluated IICA. The results shows the feasibility and robustness of the approach.

keywords: Ontology, Internet, WWW, information gathering, information classification, information exstraction and reorganization

1　はじめに

最近, 広域ネットワーク上の情報氾濫に対応するため, 様々な分野で研究が行なわれている [1], [2]. しかし, 文献検索やフルテキストサーチに基づく従来の情報検索システムでは, 大量の検索結果が整理されないまま出力されることが頻繁に発生するため, 整理・理解の支援を実現するには至っていない.

我々は, 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収集・分類・整理する IICA (Intelligent Information Collector and Analyzer) と呼ぶシステムを作成してきた [6].

本稿では, IICA の理解支援の能力をより強化するために, 新たに追加した内容抽出・統合化機能の実現方法について詳しく説明する. これは, 対象分野に特有の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, 該当する記述部分を自動抽出・統合化して提示する方法である.

さらに, IICA の収集・分類・抽出の各機能について, 人工知能に関する英文の WWW ページ, 観光に関する和文 WWW ページを対象に詳しい実験的評価を行なった (ただし, 抽出実験は後者のみ行なった). その結果, オントロジーおよび簡単なヒューリスティックスを用いた我々の方法が実用的なシステムにも適用可能であることが明らかになった.

2　知的情報収集・分類システム:IICA

本章では, 情報抽出・統合化機能を追加した IICA の概要について簡単に説明する. 図 1に IICA の概要を示す.

このシステムは, (1) 情報収集: ユーザからのキーワード列入力に応じて, WWW 上を情報を自動的に探索・収集する. このとき, オントロジーを利用して, ユーザからの依頼と関連性の高い項目は何であるか推論を行ない, 必要で思われる情報を収集する. 次に, (2) 情報分類: 収集した情報群をオントロジーに結び付けることで, 体系的に分類・整理する. さらに (3) 内容抽出・統合化: オントロジーのクラスごとに定義されている, キーワードやフレーズに着目した情報抽出ルールを用いて, 収集した各 WWW ページから該当する記述部分を自動抽出し, 統合化した結果を出力表示する. 図 2は, 温泉に関して収集した複数の WWW ページから, 温泉名, 最寄り駅, アクセス方法, 風呂の種類, 泉質といった情報を抽出・統合化した例である.

次章では (3) の方法について詳しく説明する. (1) および (2) に関する技術的な詳細は [6] を参照のこと.

3　テキストからの情報抽出・統合化

本章では, 簡単なヒューリスティックスを利用したテキストからの情報抽出・統合化法について述べる.

3.1　ヒューリスティックスに基づくテキストの内容処理

テキストの内容処理に関する最近の研究としては, 会告記事にみられるスタイル上の特長や言語表現パターンのみを利し, 電子ニュースから会議情報のダイジェスト自動生成を試みた佐藤ら [4] の報告や文章のまとまりや文の間の修辞構造に基づく抄録の自動生成を試みた住田ら [3] の報告などが挙げられる.

また, 松尾ら [7] は, 金属材料論文特有の言語表現パターンと KP と呼ばれる 技術情報の抽出法と構造化法を一体化したドメイン知識のパッケージによって金属材料論文の要約・比較を行なう METIS システムを開発している.

これらの研究は, 言語表現パターンや文書の構造などに基

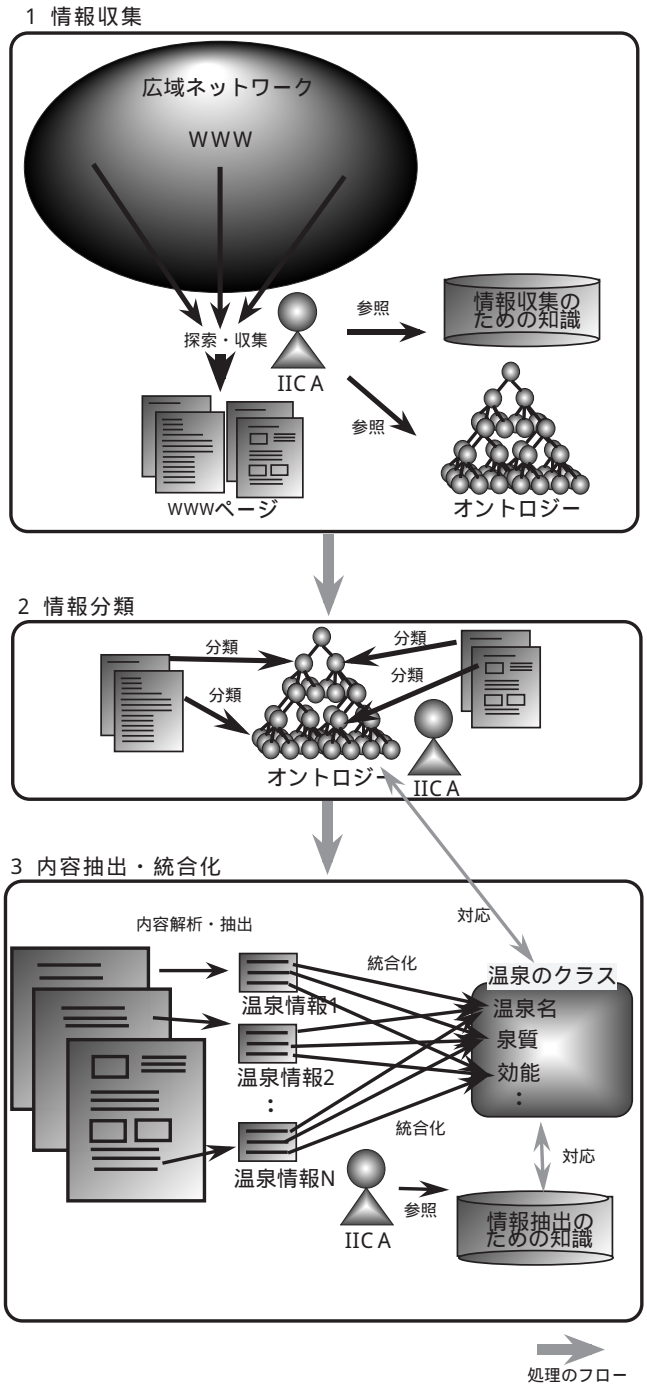

Fig 1: IICA の概要

URL	温泉の名前	最寄り駅	アクセス方法	風呂の種類	泉質
akase-spa-j.html	"赤瀬温泉"		"バス"		"炭酸鉄泉"
hinagu-spa-j.html	"日奈久温泉"	"JR八代駅"	"JR日奈久駅下車"		"食塩泉" "単純"
kanaketa-spa-j.html	"金桁温泉"	"JR三角駅"	"バス"		"炭酸鉄泉"
tsurugiyama-spa-j.html	"鶴木山温泉"	"JR佐敷駅"			"単純"
tsuruyu-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"
yunoko-spa-j.html	"児温泉"	"JR水俣駅"	"バス"	"沖合いの湯"	"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"

Fig 2: 温泉情報の統合化例

該当するHTMLファイルを(1),(2),(3)の手順で解析

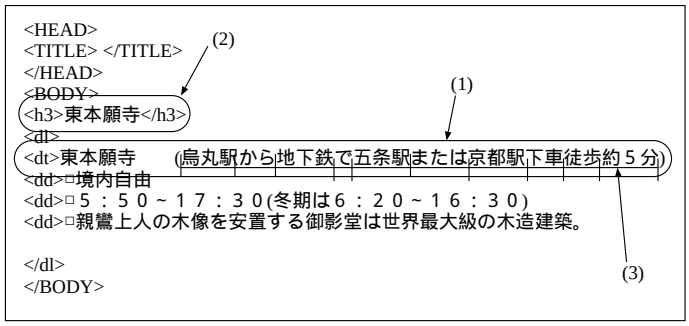

Fig 3: 交通手段情報の獲得過程

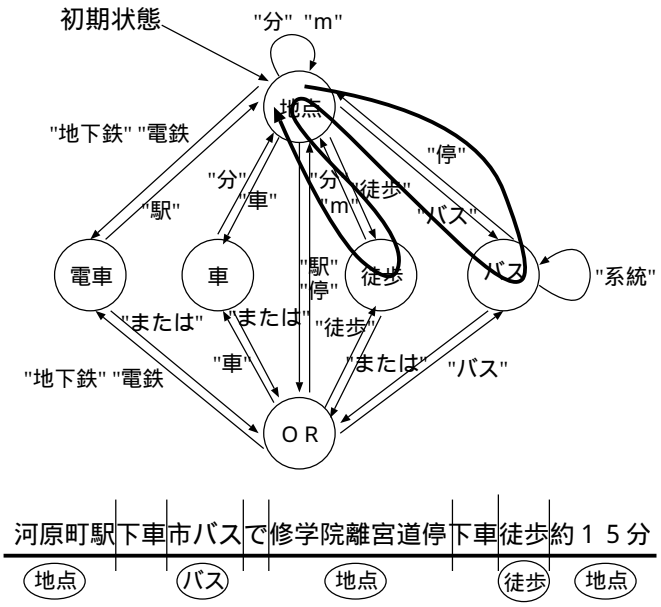

Fig 4: 交通手段解析の状態遷移図

づいた比較的浅い自然言語処理によって情報抽出・統合化を行なっている点で, 本研究と類似している. しかし, いずれのケースも対象がかなり限定されており, 他の分野への適用可能性については疑問が残る.

我々は, 例題として日本国内の観光情報が公開されているページを対象に分析し, 次の 2 つの方法ような比較的浅い自然言語処理を用いて, 実用に耐え得る情報の抽出・統合化が可能であるという確証を得た.

1. 状態遷移図を用いた方法
   これは, 状態遷移図にしたがって, 文章から特定の情報のみを順番に抽出する方法である. 例えば, 交通手段に関する情報の場合, 地点 → バス → 地点 → 徒歩 → ...,
   といった順序関係を状態遷移図によって把握しながら, 次に抽出べき情報を決定する.
2. 概念の記述ルールを用いた方法
   オントロジーの各概念に対して, 抽出すべき属性情報を定義し, 各属性情報に特有の言語表現パターンに基づたルールによって内容抽出を実行する方法. WWW 上で公開されてる, 各種の情報に適用可能である.

これらの方法は, 情報集出を単純なルールの組合せで表現する点, オントロジーにルールを結びつけることで, 構成的にルールを作成できる点を特徴としてもつもので, 多様な領域の情報源に容易に適用可能である.

以下では, 各方法についてより詳しく説明する.

3.2 状態遷移図を用いた方法

WWW ページに記述されている交通手段情報を例題に, 状態遷移図による内容の自動抽出法について述べる.

3.2.1 基本アルゴリズム

状態遷移図による情報抽出の処理は, 次の 3 つのプロセスで構成される (図 3参照).

(1) 交通手段に関する情報の発見
(2) 交通手段が書かれている観光地名の取得
(3) 交通手段の記述部分の解析と抽出

ここでは, (3) の処理について詳しく説明する.

3.2.2 交通手段の記述部分の解析と抽出

交通手段の記述部分の解析と抽出方法について, 具体的に説明する. 図 4は, ある寺院に関するページ内に記述されている交通情報の解析・抽出の原理を示したものである. この図の上側に示されているのが, 解析に用いる状態遷移図である. また, その下の文章が解析対象となった交通手段の記述部分である.

解析処理は, 状態遷移図に書かれている太い矢印の順番で行なわれる. まず, 初期状態から交通手段の記述部分の中から起点となる駅名を取得する. 次に, その駅名から状態遷移図に従い, 「バス」という単語が発見されると,“バス” の状態に遷移し, その情報を取得する. 次に, 「停」という文字が現れたら, バス停に関する記述と判断し, 再び “地点” の状態に遷移し, バス停名を取得する. 以下同様な処理を繰り

```
(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)
                    )))
```

Fig 5: 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>")))
```

Fig 6: 温泉の効能抽出のためのルール

返し, 記述部分の終端まで到達すれば, 終了し, 解析・抽出結果を出力する.

3.3 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.

3.3.1 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 5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.」), *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 6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, “+” はワイルドカードを表す記号で, “+ 症” の場合は, “冷え症”, “高血圧症” などの表現にマッチする.

Table 1: AI ページに関する精度の評価

探索法	○ (%)	△ (%)	× (%)
1 幅優先探索	64.6	7.4	28.0
2 オントロジー	66.6	11.6	21.8
3 オントロジー + 常識	67.8	10.6	21.6

Table 2: 観光ページに関する精度の評価

探索法	○ (%)	△ (%)	× (%)
1 幅優先探索	57.4	8.4	34.2
2 オントロジー	59.5	15.6	24.9
3 オントロジー + 常識	59.5	15.6	24.9

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 6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.

4 評価実験と考察

ここでは, IICA の情報収集・分類・抽出機能についての評価実験の結果と考察を行なう.

4.1 収集に関する評価

オントロジーを用いた報収集法について, WWW を対象にした実験に基づき評価を行う. 実験は, オントロジーおよび常識の利用によって, 収集精度および収集効率の 2 つの観点から行った.

4.1.1 収集の精度に関する検証

収集の精度を調べるために, 収集するページ数を 100 件に制限して, AI(英語) および観光 (日本語) に関する 5 種類の問合せに対して, 次の 3 種類の方法でそれぞれ収集実験を行なった.

a. 幅優先探索知識は全く使用しない, ブラインドサーチである. ページのアンカーを幅優先探索でたどり, 入力キーワードが含まれていればそのページを収集する.

Table 3: 収集効率の評価 — 1 キーワード： “knowledge base”

探索法	○	△	×
1. 幅優先探索	3	3	3
2. オントロジー	21	8	12
3. オントロジー + 常識	44	13	25

Table 4: 収集効率の評価 — 2 キーワード： (“semantic network” AND “production system”)

探索法	○	△	×
1. 幅優先探索	0	0	0
2. オントロジー	10	12	11
3. オントロジー + 常識	18	23	15

b. オントロジーの利用
探索では, オントロジーによるアンカーのフィルタリングを行ない, 入力キーワードまたは関連語が含まれていればそのページを収集する.

c. オントロジーと常識の併用
探索では, オントロジーの他に常識によるアンカーのフィルタリングを行ない, 入力キーワードまたは関連語が含まれていればそのページを収集する. ここで, 常識とは「人工知能の情報は研究所に多い」といった知識を連想形式で記述したヒューリスティックスである. 詳細は, [5] を参照.

収集したページの評価は, 次の 3 段階の基準に従って行なった. 5 つの問合せに対する評価の平均値を, 表 1および表 2に示す.

幅優先探索とオントロジーの利用する方法ではヒット率に若干の差があった. 特に, △のグループに属するページには明らかな差が見られ, オントロジーを用いる効果が認められた. 常識については, 精度に関する影響は確認できなかった.

○: 問合せに該当するページ.
△: 問合せの内容と異なるが, 関連性があるページ.
×: 関連性のないページ.

4.1.2 収集効率に関する検証

収集効率を調べるために, 訪問するページ数 (探索ステップ数) を 500 に固定し, AI(英語) に関する 2 種類の問合せに対して, 前述の 3 種類の方法でそれぞれ収集実験を行なった. 表 3は, 1 つキーワード (“knowledge base”) からなる問い合わせに対して収集した結果, 表 4は, 2 つキーワードの AND 条件 (“semantic network” AND “production system”) からなる問い合わせに対して収集した結果である. これらの表より, 1, 2 および 3 の方法で収集効率に明らかな違いがあることが分かる. 特に, 2 キーワードの実験 (表 4参照) では, 1 の方法で全く該当ページが収集できなかったの対し, オントロジーおよび常識を併用した方法では少数ではあるが該当するペ—ジ収集することができ, その効果が顕著に現れたといえる.

Table 5: WWW ページ分類実験の評価

	AI のページ (英語)	観光のページ (日本語)
適合率	81.9	79.0
再現率	80.5	70.0

以上の結果から, オントロジーの使用によって, 収集精度が数%向上することが分かった. 特に, 関連情報の収集に関しては約 2 倍の効果があった. また, オントロジーは収集効率にも効果をもたらすことが分かった. さらに, 常識として数個の簡単なヒューリスティックスの併用することで収集効率に約 2 倍の差があることが分かった.

4.2 分類に関する評価

4.2.1 WWW ページの分類評価

AI (knowledge base) に関して収集した約 500 件の WWW ページおよび, 観光全般に関して収集した約 800 件の WWW ページに対して, 3 章の方法に従って分類実験を行なった.

評価は, 以下の式を用いて, 再現率 (R:Recall) および適合率 (または精度) (P:Precision) を求めた. 表 5に結果を示す.

$$R = \frac{\text{正しくカテゴリに割り当てられたテキスト数}}{\text{そのカテゴリに割り当てられるべきテキスト数}},$$

$$P = \frac{\text{正しくそのカテゴリに割り当てられたテキスト数}}{\text{そのカテゴリに割り当てられたテキストの数}}.$$

4.2.2 考察

英文の AI に関するページ, 和文の観光に関するページで約 8 割程度の適合率が得られた, 再現率に関しては, 和文の観光関連ページの方が 70% と若干精度が落ちている. この原因としては日本語の形態素解析が不十分であることがあげられる. 誤って分類されたページを分析した結果, 例えば, 「店」というカテゴリーに分類されるべきテキストが,「施設」に誤って分類されてたり,「料理」に誤って分類されてしまうものが頻繁に存在することが分かった. 前者の誤りについては関連語を修正すればある程度改善できると思われる. 後者の誤りについては, 店に関するページの多くが飲食関係のものであるため, 名詞辞書の整備によって, さらに精度が向上すると思われる.

4.3 内容抽出に関する評価

ここでは, 4 章で提案した 2 つ情報抽出法を用いて, 観光に関するページ (日本語) を対象に行った評価実験の結果に

Table 6: 状態遷移図による内容抽出法の評価

1. 正確に記述部分を発見したページの割合	85 %
2. 正確に記述部分を解析・抽出したページの割合	70 %

Table 7: 概念の記述ルールによる情報の抽出実験結果

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 %

ついて示す.

4.3.1 交通情報抽出評価実験

状態遷移図に基づく抽出法を評価するために, 交通手段の書かれている WWW ページ 100 件を対象に, 以下の手順で実験を行なった. 実験結果を表 6に示す.

1. それぞれのデータについて交通手段の記述部分と観光地を抽出する.
2. その抽出した部分を解析し, その結果を出力する.
3. 抽出した部分と解析結果を評価.

1 は対象とする全ページと交通手段か記述されている部分を正確に発見し抜きだしたページの割合を, 2 は対象とする全ページから交通手段に関する情報を状態遷移図によって解析・抽出したページの割合を示している.

4.3.2 考察

1 および 2 の結果から, 記述部分の発見が正しく行なわれれば, 80 % 以上の精度で解析・抽出が可能であること分かった. 解析・抽出に関しては, 状態遷移図のより詳細な記述, 形態素解析の精度向上によって更に正確に処理が可能と思われる.

4.3.3 概念記述ルールによる内容抽出法の評価

温泉, 飲食店, お寺に関するページ各 100 件に対して, 情報抽出実験を行なった. 評価は, 分類実験の場合と同様に, 再現率および適合率を次式を用いて求めた. その結果を表 7に示す.

再現率 = 正しく取れた$item$数/本来取るべき$item$数
適合率 = 正しく取れた$item$数/実際取れた$item$数

4.3.4 考察

適合率が約 8 割, 再現率が約 6 割であった. 抽出できた情報については精度が高いことが分かった. 表 7では, 飲食店に関する再現率が他と比べ精度が極端に低いことが分かる. これは, 飲食店のページに出現する言語表現パターンが多様であるため, 予め用意したルール内の辞書だけで十分対応できなかったためである. 食品に関する専門用語辞書が整備できれば, 実用レベルまで精度を改善可能である思われる.

5 おわりに

本研究では, オントロジーおよび言語表現パターンに基づくヒューリスティックスを用いた情報抽出・統合化法を提案し, IICA の新しい機能実現として実現した. また, WWW ページを対象に IICA の各機能の評価実験を行なった. 今回の実験から, 我々の方法には, (1) オントロジーおよび常識の利用によって情報の収集精度・効率が向上する. (2) オントロジーの階層関係をたどることによって, 隣接する概念のカテゴリ—に誤って分類された情報も検索可能である, (3) 言語表現パターンに基づく単純なヒューリスティックよって, テキストの内容抽出・統合化が容易に実現可能である, などメリットがあることが明らかになった. 現在のシステムの課題は, (1) 概念間の重みは変更できるがオントロジーの構造そのものが固定的であるため, ユーザの興味や新しい情報に柔軟に対応できない, (2) 利用可能なオントロジーの数が少ない, といった点が挙げられる. これらの問題に対処するために, 収集したデータから, 新しいリンクの生成を行ない, オントロジーをユーザに適応化させる方法, 新しい概念を学習する方法などを検討している.

参考文献

[1] P. Maes and R. Kozierok. Learning interface agents. In *Proceedings of AAAI*, 1993.

[2] O McBryan. Genvl and wwww: Tools for taming the web. In *Proc. 1st Int. WWW Conf.*, 1994.

[3] 住田一男, 知野哲朗, 小野顕司, 三池誠司. 文書構造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, number 3, pp. 511–519, 1995.

[4] 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, volume 36, pp. 2371–2379, 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. 情処研報 95-FI-40, volume 95, pp. 25–32, 1995.

[6] 岩爪道昭, 武田英明, 西田豊明. オントロジーを用いた情報の自動収集と分類へのアプローチ. 人工知能学会全国大会 (第 9 回) 論文集, pp. 387–390, 1995.

[7] 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か関する研究. 信学技報 AI95-33, volume 95, pp. 87–94, 1995.