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・分類・統合化

岩爪道昭[†] 白神謙吾[†,☆] 畑谷和右[†,☆☆]
武田英明[†] 西田豊明[†]

本稿ではオントロジーに基づく情報の収集・分類・統合化法を提案する．近年，インターネットに代表される情報ネットワークで提供される情報源は急速に多様化，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．従来の情報検索ツールには，対象領域に関する体系的な知識が欠如しているため，ユーザの要求と関連のあるものは何であるかを判断したり，検索結果を体系的に分類して分かりやすく示すことは不可能であった．我々は，情報ネットワークに散在する大量の情報群から，オントロジーおよび対象領域特有の言語表現パターンに基づくヒューリスティックスを利用して，必要な情報を自動収集・分類・統合化するIICAを考案した．また，WWWを対象した評価実験の結果，我々のアプローチが，広域ネットワークに散在する大量の情報の理解支援に有効であることが分かった．

Ontology-based Information Gathering, Categorization and Reorganization from Wide-area Networks

MICHIAKI IWAZUME,[†] KENGO SHIRAKAMI,[†,☆] KAZUAKI HATADANI,[†,☆☆]
HIDEAKI TAKEDA[†] and TOYOAKI NISHIDA[†]

In this paper, we propose a new method of gathering, categorizing and reorganizing information using ontologies and heuristics. The number and diversity of information resources on the Internet is increasing rapidly. As more information become available on the Internet, it becomes increasingly difficult to acquire knowledge we need. Although many tools are available to help people to search for information they need, they cannot interpret the result of their search due to lack of knowledge. We implemented a system called IICA(Intelligent Information Collector and Analyzer) which gathers, categorizes and reorganizes information from distributed wide-area networks using ontologies and heuristics. We tested IICA for tasks on the WWW. The result of the experiments indicated that the our approach enable us to use heterogenious and very large information resources on wide-area networks.

1. はじめに

近年，インターネットに代表される広域情報環境の整備やWWW（World Wide Web）などのマルチメディア情報技術の急速な進歩・普及は，そこで提供される情報の多様化・複雑化・大規模化をもたらした．すでに，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超えている．我々がネットワークから必要な情報や知識を得るためには，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，様々な分野で研究が行われている．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．ユーザが欲しい情報だけを，効率良く収集し，理解しやすい形に分類・整理できるようなシステムが必要である．

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・統合化するIICA（Intelligent Information Collector and Analyzer）と

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
☆ 現在，三菱電気コントロールソフトウェア株式会社
Presently with Mitsubishi Electric Control Software Company
☆☆ 現在，松下電器株式会社
Presently with Matsushita Electric Industrial Co., Ltd.

呼ぶシステムを作成した．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を推論し，必要と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によって，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2参照）．

WWWを対象とした評価実験の結果，我々のアプローチが，広域ネットワークの多様な情報源の利用に有効であることが明らかになった．

2章では，IICAがネットワークからの情報収集にオントロジーをどのように利用するのか示す．3章では，オントロジーに基づく情報の分類法について説明する．4章では，簡単なヒューリティックスを利用したテキストからの情報抽出・統合化法について説明する．5章では，2〜4章で提案した各方法ついて，WWWを対象とした実験結果をもとに評価を行う．6章では，関連研究と本研究との比較および議論を行う．

2. オントロジーに用いた情報収集

本章では，オントロジーを利用した，WWWにおける情報収集法について説明する．

2.1 オントロジーについて

オントロジーは，概念化の仕様を記述したものである[1]．一般に，オントロジーの記述は，フレーム型言語や一階述語に基づく知識表現言語などが用いられる．しかし，これらの言語を用いて，まったく何もない状態からトップダウン的に大規模なオントロジーを構築することは，多大な時間と労力を要し現実的ではない．また，実世界の情報は矛盾を多く含んでおり，あらかじめすべてを考慮して体系的，網羅的に記述することはきわめて困難である．

そこで，本アプローチでは，形式的な操作性は失われるが，より現実のデータに対応するため，既存の概念体系や専門用語シソーラス，辞書などから，概念を表す語彙の集合と概念間の連想的な関係のみを記述した**弱構造化オントロジー**を採用する．

弱構造化オントロジーの構築は，まず既存の概念体系をオントロジーの雛型としてシステムに適用し，その結果から不足の概念や属性を追加しながら，現実の情報に対応するオントロジーを手動で作成した．

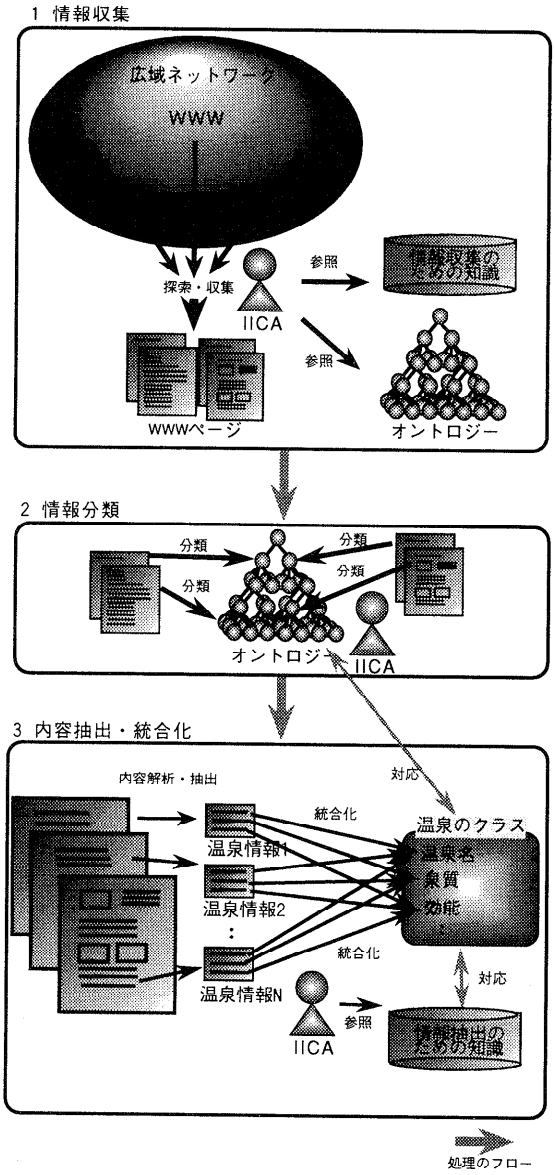


図1 IICAの概要
Fig. 1 Outline of IICA.

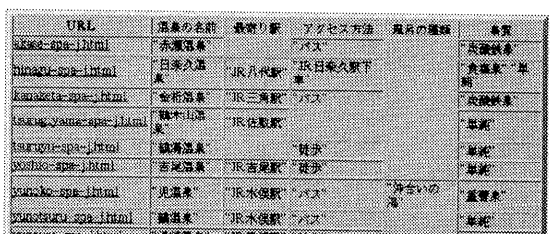


図2 温泉情報の統合化例
Fig. 2 An example of reorganization of hot-spring information on the WWW.

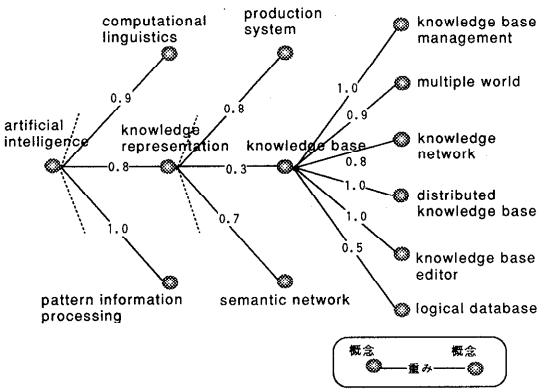


図3 弱構造化オントロジー
Fig. 3 A weak-structured ontology.

本研究では，以上の方法に基づいて，AIに関するオントロジーを情報科学辞典[2]の分野構成をもとに，観光に関するオントロジーは旅行ガイドブックやパンフレットをもとにを作成し，評価実験を行っている．**図3**は，AIに関する弱構造化オントロジーの一部を示している．図中の各ノードおよびアークはそれぞれ概念および概念間の連想関係を表している．

情報収集におけるオントロジーの役割は，システムがユーザからの問合せに関連する情報を推論し，探索空間を絞り込むための知識を提供することである．次節以降では，情報収集におけるオントロジーの利用法について詳しく述べる．

2.2 収集アルゴリズム

探索は基本的に幅優先探索で行う☆．探索の実行には，キーワード，スコープパラメータ，収集ページの3種類の入力を必要とする．システムはキーワード，スコープパラメータから，ユーザの要求と関連性のある語彙をオントロジーからリストアップし，評価値を与える．評価値を与えられた関連語のリストを利用して，次にアクセス・収集するページを決定する．以下にそのアルゴリズムの概要を示す（**図4**参照）．

step 1
求める情報に関するキーワード列，探索開始点となるURLアドレス，スコープパラメータ，収集するページ数の入力．

step 2
指定されたスコープパラメータの範囲内で，オントロジーを用いて入力キーワードの関連語をリストアップする．

step 3
指定されたURLアドレスをすでに収集しているかどうか調べる．新規のURLの場合はHTTPにアクセスしてページを収集する．

step 4
収集したページが指定された数に達していれば処理終了．満たなければstep 5へ．

step 5
収集したページをパージングし，タイトル，アンカに記述されているURLアドレス，ラベルを抽出．各URLアドレスがすでに収集済かopen-listの要素の場合は削除．それ以外は，step 6へ．

step 6
タイトルおよびラベルの中にstep 2でリストアップされた関連語が含まれているか調べる．含まれているキーワードの重みからハイパーリンクに評価値を与える．open-listに加え，評価値に従ってソートする．

step 7
アンカが収集ページに存在しない場合には，open-listからURLアドレスを取り出す．step 3へ．

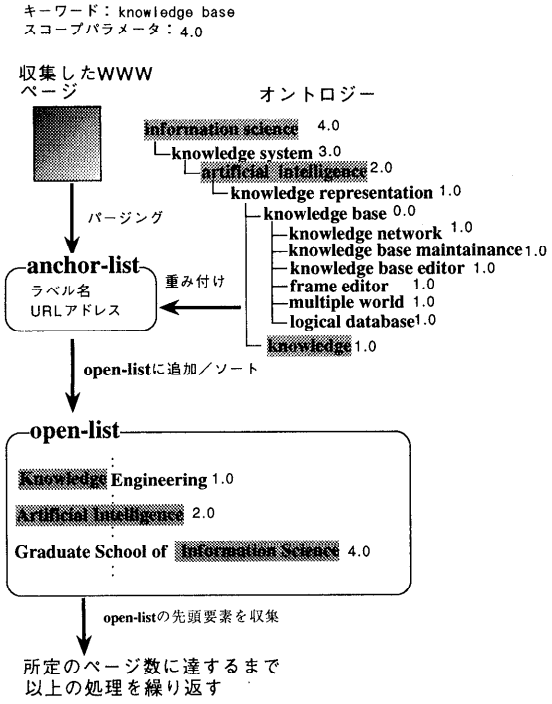


図4 WWWにおける情報収集の例
Fig. 4 An example of information gathering on the WWW.

☆ ロボットによるWWWのオフラインサーチはネットワークに非常な負担をかけるので，ユーザは細心の注意が必要である．実際の利用では，上記のアルゴリズム以外に，ネットワークへのアクセス頻度や時間制限を行ったり，特定のホストに集中してアクセスしない，といった対処が必要である．

2.2.1 例

上記のアルゴリズムについて，具体例を用いて説明する．ユーザからの入力がキーワード “knowledge base”，スコープパラメータが 4.0 と仮定する（図 4 参照）．IICA はまず，図 4 の右上にあるように，ユーザが入力したキーワードの関連語をリストアップする．例では，スコープパラメータが 4.0 であるため，“knowledge base” からの距離が 4.0 以内にある語彙がその距離とともにリストアップされる．そして，抽出したアンカのラベルにこれらの関連語が含まれていた場合には，そのアンカに評価値として，その関連語と入力キーワードとの距離と同じ値を与える．たとえば，関連語 “knowlede” とキーワード “knowledge base” との距離は 1.0 なので，“knowledge” を文字列として含むラベルを持つアンカの評価値は 1.0 となる．複数の関連語を含む場合は，そのうちの最も良い（小さい）値を与える．

2.3 常識の利用

前節で述べたアルゴリズムは，オントロジーによってアンカをフィルタリングし，探索空間を絞り込むことがねらいであった．

一方，我々が WWW 上の情報を探す場合，対象領域の知識だけでなく，常識や経験的な知識といった様々なヒューリスティックスを用いて，どのリンクをたどるか判断している．たとえば，人工知能に関する情報を探す場合には，「人工知能に関するページは大学・研究機関に多い」といった知識を利用して，大学や研究機関のページを優先的に調べるほうが，人工知能の関する情報にたどり着く可能性が高い．そこで，対象領域に関する常識の利用を試みる．

ここでは，このような常識を連想関係によって記述する．たとえば，「人工知能に関するページは研究所を探す」という常識は，

```
‘‘artificial intelligence’’
  → ‘‘laboratory’’
```

などの簡単なヒューリスティックスして与える．実際の処理では，ユーザの問合せに “artificial intelligence” およびその関連語が含まれている場合に，“laboratory” というキーワードをタイトルに含んでいるページ内のアンカを優先的に探索するように，重みを変更している．

3. オントロジーに基づく情報分類

本章では，オントロジーを利用した情報の分類法について説明する．

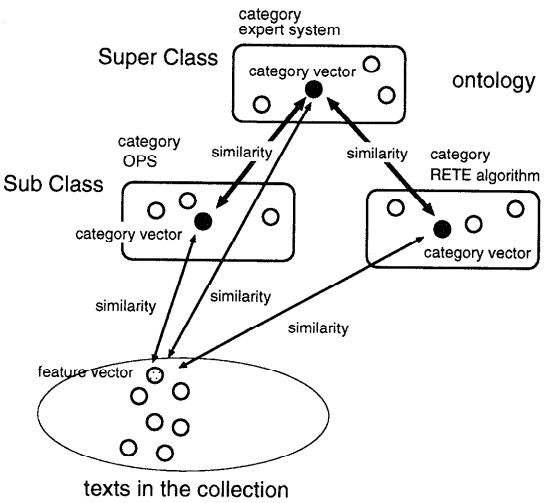


図 5 オントロジーによるテキスト分類
Fig. 5 Outline of text categorization using an ontology.

3.1 基本アルゴリズム

本研究では，オントロジーの各概念を分類のためのカテゴリとし，各カテゴリに収集したページを割り付ける方法をとる．

オントロジーに基づく分類は，(1) 特徴ベクトルとカテゴリベクトル間の類似度計算によるテキストの割り付け，(2) カテゴリベクトル間の類似度からオントロジーの概念間の重みを変更する，という一連のプロセスによって構成される（**図 5** 参照）．

ここで，特徴ベクトルとは，そのドキュメントの特徴を表すベクトルであり，後述するベクトル空間モデルに基づいて計算する．カテゴリベクトルは，そのカテゴリの特徴を表すベクトルで，そのカテゴリに分類されたテキストの代表ベクトルの重心を計算することで求められる．

(1) では，オントロジーの各概念の出現頻度を要素とするキーワードベクトルを用いて，ページとカテゴリとの類似度を計算し，類似度が最大またはあらかじめ設定した閾値以上になるカテゴリページを割り付ける．(2) のプロセスでは，オントロジーの概念間の重みを修正することで，現実のデータへの柔軟な対応を可能にする．

3.2 カテゴリベクトルの初期値計算

カテゴリベクトルの初期値計算の手順を以下に示す．

step 1 収集したテキストの特徴ベクトルの計算．

step 2 各カテゴリの代表特徴ベクトルを決定するために，収集したデータを単純なキーワードマッチングで分類する．

step 3 分類されたテキスト群から各カテゴリの特

徴ベクトルを計算.

step 4 計算した特徴ベクトルに基づいて収集テキストの再分類を行う.

step 5 各カテゴリの特徴ベクトルが収束するまで step 3, step 4 を繰り返す.

3.3 ベクトル空間モデル

本アプローチでは，単語の重み付けと，収集したドキュメントの特徴ベクトルを計算するために，情報検索の分野で広く利用されているベクトル空間モデル[3]を採用している.

単語の重み付けは，出現するテキストにおけるその単語の相対出現頻度 tf（term frequency）とテキストの集合におけるその単語の逆文献頻度 idf（inverse term frequency）の積によって与えられる．すなわち，

$$w_{ik} = tf_{ik} \times idf_k$$

ここで，tf_{ik} はドキュメント i における語 t_k の出現頻度，idf_k はドキュメント集合において語 t_k が出現したドキュメントの数の逆数である．一般に用いられる idf の尺度は次式で与えられる.

$$idf_k = \log(N/n_k)$$

ここで，N はドキュメントの総数であり，n_k はキーワード t_k を含むテキストの数である.

4. 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.

例として，日本国内の観光情報が公開されているページを収集・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．たとえば，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 上で公開されている，各種の情報に適用可能である.

以下では，各方法についてより詳しく説明する.

4.1 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

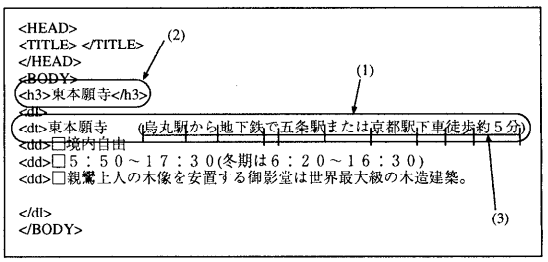


図 6 交通手段情報の抽出過程

Fig. 6 A process of extracting information about means of transportation.

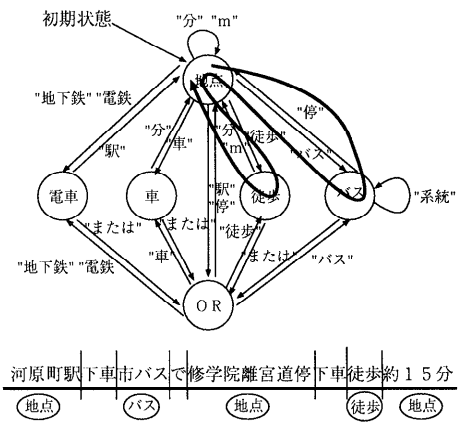


図 7 交通手段解析の状態遷移図

Fig. 7 A state diagram for analysis of information about means of transportation.

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.

4.1.1 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 6** 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.

4.1.2 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 7** 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.

4.2 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.

4.2.1 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だめのルール設定は,次の手順で行う.

(1) 属性情報の設定

たとえば,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8**に定義記述例を示す.この例では,「**温泉**は,値を1つとる**温泉の名前**という属性と,値を1つ以上とる**風呂の種類**,**泉質**,**効能**という属性を持つ**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うルール部分の設定を行う.**図9**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,“+”は長さが1以上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,手作業で行っている.ルールの設定自体は,どの概念でも図9のような形式で統一的に記述可能で,属性の数は多い場合で十数個程度,各属性の情報抽出ルールも長いもので数行程度である.

```
(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8 温泉に関する属性情報の定義
Fig. 8 A definition of extraction items.

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

図9 温泉の効能抽出のためのルール
Fig. 9 Attribute extraction rules for hot-springs.

5. 評価

ここでは,2~4章で提案した情報収集・分類・抽出の各方法について評価し,その有効性について検証する.

5.1 収集に関する評価

オントロジーを用いた情報収集法について,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.

5.1.1 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,収集実験を行った.

a. 幅優先探索

ページのアンカ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.知識はまったく使用しない.

b. オントロジーの利用

幅優先探索の際,オントロジーによるアンカのフィルタリングを行い,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.

表1 AIページに関する精度の評価
Table 1 Evaluation of gathering pages relevant to artificial intelligence.

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

表2 観光ページに関する精度の評価
Table 2 Evaluation of gathering pages relevant to sghitseeing.

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

c. オントロジー+常識の利用

幅優先探索の際，オントロジーおよび常識によるアンカのフィルタリングを行い，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．

収集したページの評価は，次の3段階の基準に従って行った．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．

5.1.2 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った．**表3**は，1つのキーワード（“knowledge base”）からなる問合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”）からなる問合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の方法でまった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ー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．

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を併用することで収集効率に約2倍の差があることが明らかになった．

表3 収集効率の評価—1キーワード：“knowledge base”
Table 3 Evaluation of efficiency of infromation gathering —1 keyword (“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

表4 収集効率の評価—2キーワード：“semantic network” AND “production system”
Table 4 Evaluation of infromation gathering—2 keywords (“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

表5 WWWページ分類実験の評価
Table 5 Evaluation of categorization of WWW pages.

	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

5.2 分類に関する評価

5.2.1 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．

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

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

5.2.2 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，たとえば，「店」というカテゴリに

表6 状態遷移図による内容抽出法の評価
Table 6 Evaluation of exatraction of traffic information using a state diagram.

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%

分類されるべきテキストが,「施設」に誤って分類されていたり,「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.

5.3 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ついて示す.

5.3.1 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, 交通手段が記述されている部分を正確に発見し抜き出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.

5.3.2 考 察

1および2の結果から, 記述部分の発見が正しく行われれば, 80%以上の精度で解析・抽出が可能であることが分かった. 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さらに正確に処理が可能と思われる.

5.3.3 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.

$$\text{再現率} = \frac{\text{正しく抽出した item 数}}{\text{本来抽出すべき item 数}}$$

$$\text{適合率} = \frac{\text{正しく抽出 item 数}}{\text{実際に抽出した item 数}}$$

表7 概念の記述ルールによる情報の抽出実験結果
Table 7 Recall and precision of extraction of information using heuristics.

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%

5.3.4 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, あらかじ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.

6. 関連研究

6.1 インターネットロボット, エージェント

最近, WWWの上の情報を探索するワーム型エージェント[4]や行動履歴などからユーザの関心事項を学習するエージェント[5),6)]の研究などが行われている. しかし, これらのシステムは対象領域に関する体系的な知識が欠如しているため, ユーザが必要とする情報がどんな分野に属するものか, 関連する情報にはどのようなものがあるのか, といったことは判断できない. また, 収集した情報を解釈して, ユーザの理解を助けることはできない. 我々のアプローチでは, システムに体系的な知識をオントロジーとして与えることで, より知的な情報収集が可能である.

6.2 情報検索, テキスト分類

シソーラスのような構造化知識を用いたテキストの分類は, すでに研究が報告されている[7),8)]. しかし, これらのアプローチでは, シソーラスが完全に固定されているため新しい情報や扱う情報の変化に対応しにくい. また, 各カテゴリの代表ベクトルが最初の学習データに大きく依存したり, カテゴリ間の関係の強さは考慮されていないため, あるカテゴリに属するテキストから意味的に近いテキストへの検索ができない, などの問題がある.

一方, Kohonenの自己組織化マップやニューラルネットワークを情報検索に適用したアプローチが近年注目されている[9),10)]. これらの方法は, 曖昧検索や組織化されたキーワードマップの可視化を利用した検索が可能などメリットも多い. しかし, データに基づく

ボトムアップなアプローチであるため，組織化されたマップの構造に意味を与えることは難しい．

本研究のアプローチは，構造化されたオントロジーの利用と収集したデータに基づいたカテゴリベクトルおよびオントロジーの概念間の重み付けの変更により，トップダウンな手法とボトムアップな手法の短所を補うことができる．

6.3 内容処理

内容処理に関する最近の研究としては，会告記事に見られるスタイル上の特長や言語表現パターンのみを利用し，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[11]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[13]の報告などがあげられる．

また，松尾ら[12]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．一方，本研究では，情報抽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付けることで，構成的にルールを作成できる点を特徴として持つもので，多様な領域の情報源に容易に適用可能である．また，状態遷移図による情報の解析・抽出法は，今までにない新しい試みである．

6.4 オントロジー

オントロジーに関する研究には，CYC[14]やSharable Ontology Library[15]のように実際に作られたものは存在する．しかし，これらの研究では，オントロジーをどのように利用するかという面に関する考察は少ない．本研究におけるオントロジーに対するアプローチは，利用面を中心に考察する点に特長がある．

7. おわりに

本研究では，オントロジーを用いた，新しい情報の収集・分類・統合化法を提案し，広域ネットワークからの情報獲得を支援するためのシステムIICAを実装した．また，WWWを対象にIICAの各方法の評価実験を行った．これらの結果から，我々のアプローチには次の5つメリットがあることが分かった．

(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) オントロジーとクラスタリングの併用によって，ユーザの目的と収集データに対応した分類が可能である．

(3) 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．

(4) 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．

(5) オントロジーを中心にシステムを構築することで，情報の収集・分類・統合化が一貫して実現できる．

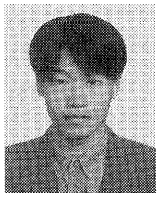
現在のシステムの課題は，(1) 概念間の重みは変更できるがオントロジーの構造そのものが固定的であるため，ユーザの興味や新しい情報に柔軟に対応できない，(2) 利用可能なオントロジーの数が少ない，といった点があげられる．これらの問題に対処するために，収集したデータから，新しいリンクの生成を行い，オントロジーをユーザに適応化させる方法，新しい概念を学習する方法などを検討している．

参考文献

1) Gruber, T.R., Tenenbaum, J.M. and Weber, J.C.: Toward a Knowledge Medium for Collaborative Product Development, *Proc. 2nd Int. Conf. on Artif. Intell. in Design*, pp.413–432 (1993).
2) 長尾 真，石田晴久，稲垣康善，田中英彦，辻井潤一，所真理雄，中田育男，米沢明憲：岩波情報科学辞典，岩波書店 (1996)
3) Salton, G and McGill, M.J.: *Introduction to Modern Information Retrieval*, MacGraw-Hill (1983).
4) McBryan, O.: GENVL and WWWW: Tools for Taming the Web, *Proc. 1st Int. WWW Conf.* (1994).
5) Maes, P.: Agents that Reduce Work and Information Overload, *CACM*, Vol.37, No.7, pp.30–40 (1994).
6) Balabanovic, M. and Shoham, Y.: "Learning Information Retrieval Agents: Experiments with Automated Web Browsing, *Proc. AAAI Spring Symposium*, pp.13–18 (1995).
7) 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol.33, No.9, pp.1114–1122 (1992).
8) 山本和英，増山 繁，内藤昭三：分類体系相互の関係を利用したテキストの自動分類，情報処理学会研究会報告，Vol.95, No.27, pp.7–12 (1995).

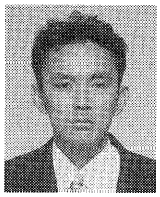
9) Kohonen, T.: The Self-Organizing Map, *Proc. IEEE*, Vol.78, No.9, pp.1464–1480 (1990).
10) 仁木和久，田中克己：ニューラルネットワーク技術の情報検索への適用，人工知能学会誌，Vol.10, No.1, pp.45–51 (1994).
11) 佐藤 円，佐藤理史，篠田陽一：電子ニュースのダイジェスト自動生成，情報処理学会論文誌，Vol.36, No.10, pp.2371–2379 (1995).
12) 松尾利行，武田英明，西田豊明：技術情報空間の構築と探訪の知的支援に関する研究，信学技報 AI95-33, Vol.95, No.265, pp.87–94 (1995).
13) 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），Vol.J78-D-II, No.3, pp.511–519 (1995).
14) Guha, R.V. and Lenat, D.B.: Cyc: A Midterm Report, *AI magazine*, pp.32–59 (1990).
15) Sharable Ontology Library: http://www-ksl.stanford.edu/knowledge-sharing/ontologies/index.html

(平成8年1月5日受付)
(平成9年1月10日採録)



岩爪 道昭（学生会員）

1968年生．1991年姫路工業大学工学部電子工学科卒業．1993年同大学工学研究科電気電子工学専攻修士課程修了．1995年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学情報科学研究科博士後期課程2年在学中．ネットワークからの情報獲得・統合，知識の共有と再利用に従事．1995年人工知能学会全国大会優秀論文賞受賞．人工知能学会，ソフトウェア科学会各会員．



白神 謙吾

1970年生．1994年京都大学工学部数理工学科卒業．1996年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．情報の分類，オントロジー獲得の研究に従事．同年三菱電機コントロールソフトウェア（株）入社．現在に至る．

畑谷 和右

1971年生．1994年京都大学工学部情報工学科卒業．1996年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．情報の抽出・獲得に従事．同年松下電器産業（株）入社．現在，流通・管理情報システムセンター勤務．

武田 英明

1986年東京大学工学部卒業．1988年同大学院工学系研究科修士課程修了．1991年同大学院工学系研究科博士課程修了．1991年（財）日本システム開発研究所嘱託研究員．1992年ノルウエー工科大学 postdoctoral fellow．1993年奈良先端科学技術大学院大学情報科学研究科助手．1995年同助教授．現在に至る．東京大学工学博士．1992年よりノルウエー王立科学技術研究会議 research fellow．ネットワークからの情報獲得と統合，分散協調型知識ベースシステム，知的CADのための設計過程のモデル化などの研究に従事．1995年人工知能学会全国大会優秀論文賞受賞．人工知能学会，AAAIなど各会員．

西田 豊明（正会員）

1977年京都大学工学部情報工学科卒業．1979年同大学院修士課程修了．同大学助手，助教授を経て，1993年より奈良先端科学技術大学院大学教授，現在に至る．京都大学工学博士．1984年から1年間 Yale 大学客員研究員．1995年から科学技術庁金属材料研究所客員研究官．知識と共有の再利用，知識メディア，定性推論の研究に従事．1988, 89, 93年人工知能学会全国大会優秀論文賞．1988年度人工知能学会論文賞受賞．1990年度情報処理学会30周年記念論文賞．著書：「自然言語処理入門」（オーム社),「定性推論の諸相」（朝倉書店）など．人工知能学会など各会員．IJCAI-97 Video Track Chair, 京都21委員など．