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

岩爪 道昭・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

本稿ではオントロジーに基づく情報の収集・分類法を提案する．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か判断したり、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、オントロジーを利用して必要な情報を自動的収集・分類する知的エージェントIICAを考案し、プロトタイプシステムの開発を行った．また、WWWおよびネットワークニュースを対象した情報収集および分類の評価実験を行った．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多様な情報の収集・分類に有効であることが明らかになった．

Ontology-Based Information Gathering and Categorization from Wide-area Networks

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

In this papaer, we propose a new method of information gathering and text categorization using ontologies. 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. Many tools are available to help people search for the information they need. However, these tools are unable to interpret the result of their search due to lack of knowledge. We need more intelligent system which facilitates personal activities of producing information such as surveying, writing papers and so on. We implemented a system called IICA(Intelligent Information Collector and Analyzer) which helps people to acquire knowledge from information resources on the wide-are network by gathering information and categorizing texts. We tested IICA for tasks on the Internt. The result of the experiments indicated that the our approach enable us to use heterogenious information resources on the wide-area network.

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などのマルチメディア情報技術の普及により, 個人でも容易にネットワーク上の情報の利用したり, 提供することが可能になっている. それにともないネットワークの利用者は急速に増加しており, そこで提供される情報源の多様化・大規模化もとどまるところを知らない. すでに, 個人で処理しなければならない情報の量も, 人間の処理能力を越えてしまったと言えよう. 我々が広域ネットワークから必要な知識を得るためには, 情報の収集・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.

このような問題に対し, 情報ネットワークのインターフェースや情報検索ツールの研究が盛んに行なわれおり, 実際に利用可能なものも存在する. しかし, これらのツールの多くは対象領域に関する体系的な知識が欠如しているために, ユーザが必要とする情報が, どの分野に属するものか理解したり, 検索結果を解釈し, 分かりやすく示すことはできない.

我々は, 特定の対象領域に関する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し, 分類・整理するIICA (Intelligent Information Collector and Analyzer) と呼ぶシステムを作成した. IICA の機能の概要を図 1に示す.

このシステムは, (1) ユーザからのキーワード列入力に応じて, www やネットワークニュース上の情報を探索・収集する. このとき, IICA は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 (図 1(a) 参照). さらに, (2) 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 (図 1(b) 参照).

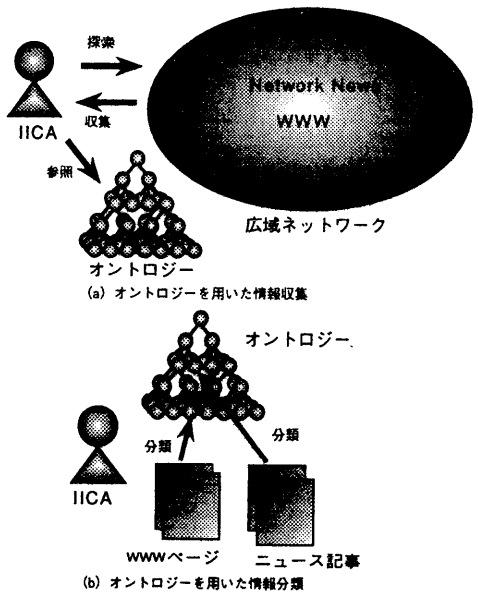

図 1: IICA:Intelligent Information Collector and Analyzer

www およびネットワークニュースを対象にした評価実験の結果, オントロジーに基づく我々の方法が, 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.

以下, 第 2 章では, 情報の収集・分類に用いる弱構造化されたオントロジーについて述べる. 第 3 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. また, www を対象にした実験結果を示す. 第 4 章では, 弱構造化オントロジーに基づく, 情報の分類法について説明する. さらに, 電子ニュースの記事および www ページを対象にした評価実験の結果について述べる. 最後に, 第 5 章で本研究で考察とまとめを行なう.

2 弱構造化オントロジー

オントロジーは, 概念化の仕様を記述したものである [1]. 情報の収集・分類において, オントロジーは, 次の 4 つの役割を果たす.

情報収集の指針
利用者が必要とする情報がどの分野に属するものか, 関連する情報は何か, エージェントが理解し, 情報収集するための指針となる.

情報フィルタ
収集してきた情報をフィルタリングによって体系的に分類・整理する.

情報検索用インデックス
分類された情報に利用者がアクセスするためのインデックスとなる.

マン・マシン共通の基本語彙の提供
マン・マシン共通の語彙を提供することによって, 人間とエージェントによる共同作業が可能になる

オントロジーは, Ontolingua[1] に代表されるフレーム型言語や一階述語に基づく知識表現言語によって記述されることが多い. しかし, これらの言語によって, トップダウン的に大規模なオントロジーを構築するには非常な時間と労力が必要であり現実的ではない. また, 実世界の情報は矛盾を多く含んでおり, 予めすべてを考慮して体系的, 網羅的に記述することは困難である. そこで我々は, 専門用語辞書や自然言語テキストなど既存の情報源からのオントロジー構築とその利用方法について研究を行なってきた [7][9]. 本研究においても同様な観点から, 既存の概念体系や専門用語シソーラスを核として, オントロジーと構築する方法をとる. このようなオントロジーの大きな特徴は, 弱構造化 [6] すなわち概念を表す語彙の集合と概念間の連想的な関係のみが記述されているということである. 図 2に, 弱構造化オントロジーの例の一部を示す. 各ノードは概念を, 各アークは概

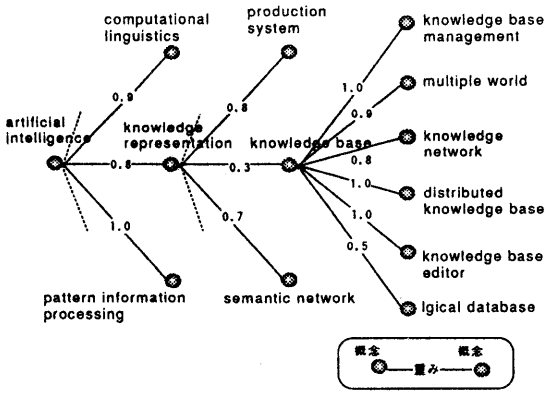

図 2: 弱構造化オントロジー

念間の連想関係を表している．アークには，概念間の結び付きの強さを表す重みが与えられているが，クラスーインスタンス，部分－全体のような概念の関係は区別しない．また，概念自体の定義も記述されていない．

本研究では，約 4,500 の情報科学の専門用語から上記のようなオントロジーを構築し，情報の収集と分類に利用している．

3 オントロジーに基づく情報収集

この章では，IICA が，オントロジーを用いて広域ネットワークからどのように情報収集するか述べる．

最近，www の上の情報を幅優先探索するワーム型エージェント [3] や行動履歴からユーザの関心事項を学習する能力を持つエージェント [2] の研究などが行なわれている．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か，といったことは理解できない．したがって，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．そこで，システムに体系的な知識を与えることで，より知的な情報収集を行なわせる方法を検討した．

3.1 WWW における情報の収集

ここでは，IICA が情報収集の際，オントロジーをどのように利用するのか，WWW における例を用いて説明する．図 3に，WWW における情報収集の概要を示す．ユーザは，キーワード，スコープパラメータ，収集ページ数からなる問合せを IICA に行なう．IICA はキーワード，スコープパラメータを用いて，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，評価値を与える．IICA は，評価値を与えられたこれらの関連語を利用して，次にアクセス・収集するページを決定する．

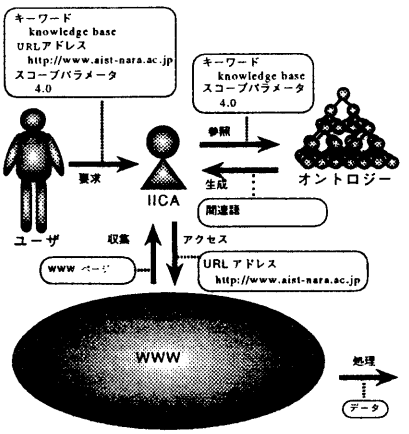

図 3: WWW における情報収集の概要

3.1.1 基本アルゴリズム

探索方法は基本的に幅優先探索であるが，収集した情報をオントロジーを用いて評価し，その次にどのページを収集するか判断する．以下にそのアルゴリズムの概要を示す．

step1

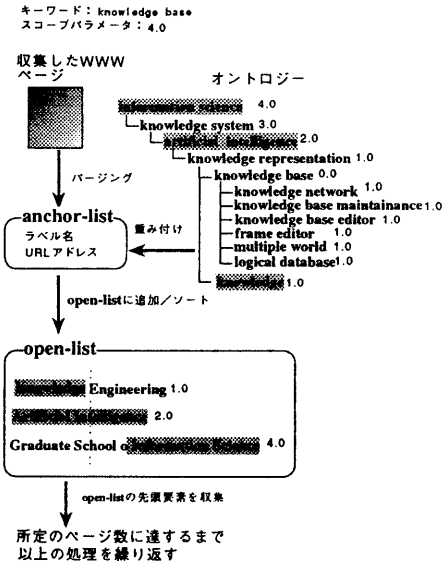

図 4: WWW における情報収集の例

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, 文脈の推定範囲, 収集するページ数の入力.

step2
キーワードとオントロジーのマッチングを取る. マッチしたオントロジー内の概念と関連性の高い情報を指定された推定範囲内でリストアップする.

step3
指定されたURLアドレスがclose-list既に存在するか（既に収集したことがあるか）調べる. 存在しない場合は, HTTPにアクセスしてページを収集する.

step4
収集したページが指定された数に達していれば処理終満たなければstep5へ.

step5
収集したページをパージングし, タイトル, ハイパーリンクのURLアドレス, ラベル名を抽出する. 各URLアドレスが既にopen-list, close-listに存在していないかチェックする. 存在していれば, open-listに追加する候補とする.

step6
タイトルおよびラベルの中にstep2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.

step7
アンカー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. step3へ.

例えば, 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であったとする (Fig 4参照). IICAはまず, Fig 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, 抽出したアンカーのラベルにこれらの関連語が含まれていた場合には, そのアンカーに評価値として、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. 例えば, 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ーの評価値は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.

3.1.2 常識の利用

オントロジーを用いた情報探索の基本的なアルゴリズムは既に述べた. この方法は, 対象領域の体系的な知識に基づいて, アンカーのフィルタリングを行なうため, 単純な幅優先探索に比べ, ユーザが求める情報をより効率良く収集できることが期待される. しかし, この方法は探索かなり強い制限を設けるため, 偶然による情報発見はあまり期待できない.

一方, 我々が, WWW上で情報収集を行なう場合, 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, どのリンクを辿るか判断している. 例えば, 人工知能に関する情報探す場合には,

「人工知能に関するページは大学・研究機関に多い.」

といった知識を用いて, 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が得られる可能性が高い. そこで, このような対象領域に関する常識の利用を試みる.

我々のアプローチでは用いる常識は, オントロジーと同様に, 連想関係のみを記述する. 例えば, 上記の人工知能の情報収集に関する常識は,

```
''artificial intelligence'' → ''university''
''artificial intelligence'' → ''institute''
''artificial intelligence'' → ''research''
''artificial intelligence'' → ''laboratory''
```

といった連想関係によって与えられる.

ユーザの問合せに, 人工知能の用語が含まれる場合には, これらの常識が適用される. 常識が適用されると, "university", "institute", "research", "laboratory" のタームがページタイトルに含まれていれば, そのページのアンカーが優先的に探索されるように重みを変更する.

3.2 評価実験と考察

これまで述べた我々の方法の有効性を検証するために, WWWにおける情報収集の実験を行なった. 実験では, 人工知能関連の内容に関する5種類の問合せに対して, 次の3種類の方法でそれぞれ100ページずつ収集した.

a. 幅優先探索
探索は幅優先探索で行ない, 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.

b. オントロジーの利用
探索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＋常識の利用
探索はオントロジーおよび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.

収集したページの評価は, 次の基準に従って手作業で行なった. 5つの問合せに対する評価の平均値を表1に示す.

○: 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.

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明らかな差があった. 特に, △のグループに属するページに差が現れており, オントロジーを用いる効果が認められた.

表 1: 評価実験の結果

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

一方, 今回の実験では, 常識をを用いた場合とそうでない場合とで, ヒット率の有意な差は認めらなかった.

また, 問合せ内容によってその評価にかなり差が見られた. "artificial intelligence" といった比較的上位の概念の問合せに対しては, オントロジーの効果が明確には現れず, いずれも高いヒット率であった (表 2参照).

一方, "reasoning" or "search" の問合せでは, "search" が人工知能以外の用語にも多く用いられる単語であるため, ヒット率は悪かった. また, オントロジーを用いた場合とそうでない場合で, 差が明確に現れた (表 3参照).

今回は, 収集した情報の内容関する評価のみで, 探索時間の定量的な評価は行なっていないが, 実効的な探索時間は, c ＜ b ＜ a の順であることを確認している.

表 2: 問合せ "artificial intelligence" の結果

探索法	○ (%)	△ (%)	× (%)
幅優先探索	86.0	9.0	5.0
オントロジー	88.0	10.0	2.0
オントロジー＋常識	90.0	9.0	1.0

表 3: 問合せ "reasoning" or "search" の結果

探索法	○ (%)	△ (%)	× (%)
幅優先探索	36.0	8.0	56.0
オントロジー	44.0	13.0	43.0
オントロジー＋常識	44.0	12.0	44.0

4 オントロジー基づく情報の分類

本研究におけるテキスト分類では, オントロジーの各概念をそれぞれカテゴリーとして設定し, 収集したテキストをこのうちのどれかのカテゴリーに割り当てる.

従来のテキスト分類では, テキストの分類の正確さだけが研究の焦点となっており, 分類の結果が理解し易いかどうか, 誤って分類された情報をどのように扱うか, といった点については考慮されていない. 我々の狙いは, オントロジーを用いて従来の手法を拡張するにより, より実用性の高い分類方法を実現することにある.

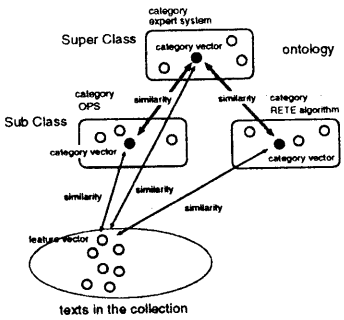

図 5: オントロジーによるテキスト分類

オントロジーに基づくテキスト分類は, (1) 特徴ベクトルとカテゴリーベクトル間の類似度計算, (2) カテゴリベクトル間の類似度計算の一連のプロセスによって行なわれる.

特徴ベクトルは, そのドキュメントの特徴を表すベクトルである. 計算方法については, 4.2 で述べる. 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, そのカテゴリーに分類されたテキストの代表ベクトルから求める.

4.1 構造化知識の利用

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 [5] [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, 各カテゴリー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, カテゴリー間の関係の強さは考慮されていないため, あるカテゴリー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.

本研究のアプローチは, 単なる構造化された分類体系によるテキスト分類法だけでなく, 収集したデータにもとづいて逐次代表ベクトルを更新するとともに, カテゴリー間の重み付けも変更することにより, 処理した現実データに対応した検索が行なえる (図 5参照). 以下に, 特徴ベクトルおよび概念間の重み (weight) の計算手順について述べる.

step1 収集したテキストの特徴ベクトルの計算.

step2 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なるキーワードマッチングで分類する.

step3 分類されたテキスト群から各カテゴリーの特徴ベクトルを計算.

step4 計算した特徴ベクトルに基づいて収集テキスト再分類を行なう.

step5 各カテゴリの特徴ベクトルが収束するまで step3, step4 を繰り返す.

step6　各カテゴリ間の距離を計算し，オントロジーの概念間の重みを変更する．

4.2　ベクトル空間モデル

我々は，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を [4] 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．

4.3　分類実験

我々は "artificial intelligence" に関する ネットワークニュースの 400 件の記事を対象に分類実験を行なった．ニュースグループ は "comp" のみ対象にした．IICA はそれらの記事を 75 のカテゴリーに分類した．表 4にその結果を示す．表の右側は上位 20 カテゴリーの記事の数，左の表は下位 20 カテゴリーの記事の数である．

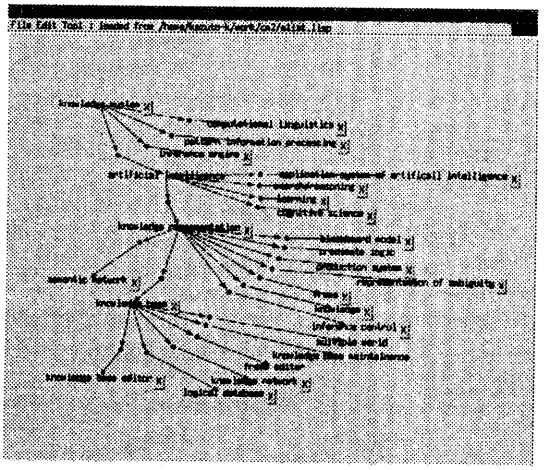

図 6: オンロトジーの可視化

表 4: 記事の分類結果

the top 20 categories and the number of texts		the low 20 categories and the number of texts	
program	48	VLSI	1
planning	31	statistics	1
aritificial intelligence	25	SQL	1
prolog	17	signal	1
software	16	psycology	1
inference engine	14	PC	1
classification	13	lisp	1
cognitive science	12	interface	1
expert system	10	informatics	1
C	9	DOS	1
Turing	8	device	1
neural network	8	design	1
TSP	7	connectionism	1
information	7	computer security	1
concept	7	compiler	1
communication	6	chess machine	1
search	6	brain	1
fuzzy	6	bag	1
IEEE	6	backpropagation	1
backtracking	6	analog computer	1

表 5: 分類実験の評価

Accuracy(%)	Recall(%)	Precision(%)
77.0	76.2	76.0

結果を評価するために，以下の式を用いて，Accuracy(A), Recall(R) および Precision(P) を求めた．

$$A = \frac{\text{正しく割り当てらたテキスト数}}{\text{収集したテキストの数}}$$

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

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

計算結果を表 5に示す，ただし，Recall と Precision はともに全カテゴリーの平均値である．

我々はまた，誤った分類結果を分析し，3 つのグループに大別した．最初のグループは，図 7(a) に示すように，正しく分類されるべきクラスの下位クラスに分類されたグループ，

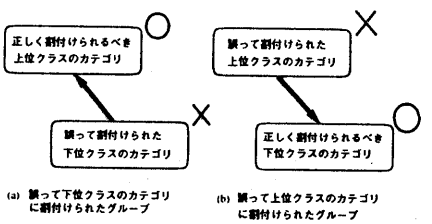

図 7: 誤り分類のグループ

2つ目グループは，図 7(b) に示すように，正しく分類されべきクラスの上位クラスに分類されたグループ，最後は，正しく分類されるべきクラスとは全く無関係のクラスに分類されたグループである．その分析結果を表 6に示す．

表 6: 誤り分類の分析結果

result type	No. of texts
下位クラスのカテゴリに割り当てられたテキスト数	26
上位クラスのカテゴリに割り当てられたテキスト数	5
上位クラスのカテゴリに割り当てられたテキスト数	51

表 7: 分類実験の再評価

Accuracy(%)	Recall(%)	Precision(%)
85.3	85.1	85.1

表 8: www ページ分類実験の結果

カテゴリー	ドキュメント数
knowledge	188
knowledge base	132
artificial intelligence	128
production system	71
knowledge representation	37
distributed knowledge base	22
blackboard model	9
semantic netwrok	5
その他 (knowledge engineering 等)	30

最初の 2 つのグループに関しては，図 6に示すようなオントロジーの構造を示すブラウザを利用して，概念間関係を辿ることで，誤って分類された情報にもアクセス可能である．そこで，この 2 つのグループを正解と仮定して再評価した結果，表 7のように分類の精度が向上した．

従来のアプローチでは，誤って分類されたテキストにはアクセスすることは困難であった．それに対し IICA では，オントロジカルな関係をたどることで誤って分類されたテキストでも，上位または下位のクラスに存在する場合には，検索が可能になることがわかった．

表 9: www ページ分類実験の評価

Accuracy(%)	Recall(%)	Precision(%)
82.1	81.9	81.9

また，我々は，問合わせ "knowledge base" によって収集された 500 個の www ページについても，上記の方法に従って分類を行なった．分類結果およびその評価を，表 8および表 9に示す．分類対象が，ドメインをかなり限定した人力に対して集められたページであったため，割り付けられたカテゴリーは 15 となった．評価については，ニュース記事とほぼ同様の結果が得られた．

以上の実験から，オントロジーを用いた我々の手法が，ニュース記事，www ページの分類に適用可能であることが明らかになった．

5 まとめ

本研究では，オントロジーを用いた，新しい情報の収集および分類方法を提案した．情報を収集・分類することによって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，我々は，WWW および電子ニュースを対象に IICA の実験を行なった．これらの実験結果から，我々の方法には次の 5 つメリットがあると言える．

- オントロジーを用いるアプローチは広域ネットワークにおける多様な情報源の利用可能にする．
- オントロジーの階層関係をたどること、誤って分類された情報にもアクセスが可能になる．
- 弱構造化オントロジーの構築は，従来の方法に比べて容易である．
- 情報の探索の際，枝刈りをおこなうのでネットワークへの負担を減らすことができる．

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を検討している．

参考文献

[1] T.R. Gruber, J.M. Tenenbaum, and J.C. Weber. Toward a knowledge medium for collaborative product development. In *Proc.2nd Int. Conf. on Artif. Intell. in Design*, pp. 413–432, 1993.

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

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

[4] G. Salton. *Introduction to Modern Information Retrieval.* MacGraw-Hill, 1983.

[5] 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. Vol. 33, No. 9, pp. 1114–1122, 1992.

[6] 花川賢治. 日常知識の弱構造化に関する研究. 奈良先端科学技術大学院大学 修士論文, 1995.

[7] 岩爪道昭, 武田英明, 西田豊明. 電子掲示版における記事の自動分類と議論の可視化－知的ニュースリーダの提案. 人工知能学会全国大会 (第 8 回) 論文集, pp. 497–500, 1994.

[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. Vol. 95, No. 27, pp. 7–12, 1995.

[9] 錦正信, 武田英明, 西田豊明. マルチエージェント系による関連知識の抽出・統合と提示. 人工知能学会全国大会 (第 8 回) 論文集, pp. 505–508, 1994.