弱構造化オントロジーを用いたインターネットからの情報獲得

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

電話: 07437-9-9211 ext.5316, fax: 07437-2-5269

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

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.

keyword: オントロジー、インターネット、情報獲得、 WWW、電子ニュース

Weakly Structured Ontology-Based Information Caputuring from the Internet

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

phone: 07437-9-9211 ext.5316, fax: 07437-2-5269

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

In this papaer, we propose a new method of information gaathering 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. A number of 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 of the domain. We need more intelligent system which facilitates personal activities of producing information such as surveying, writing papers end 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 heterogenuous information resources on the wide-area network.

keyword: Ontology, Internet, Information Capturing, WWW, Network News

1　はじめに

本稿では、オントロジーを用いた情報の収集と分類法を提案する.

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.

しかし、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. ネットワーク上の情報を利用してサーベイや論文の作成を行なう、といった情報産活動を支援するためには, より知的なシステムが必要である.

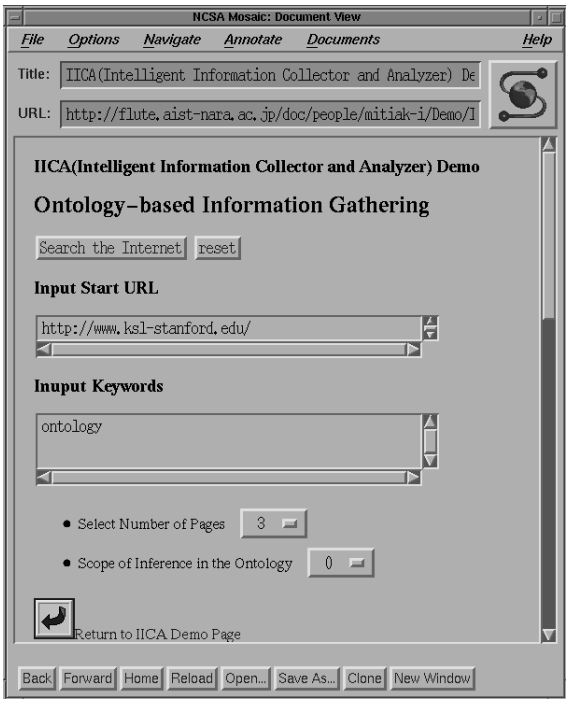

Fig 1: IICA の Mosaic インターフェース

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 (Fig 1参照). IICA の機能の概要を Fig 2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 (Fig 2(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 (Fig 2(b) 参照).

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 章で本研究の考察とまとめを行なう.

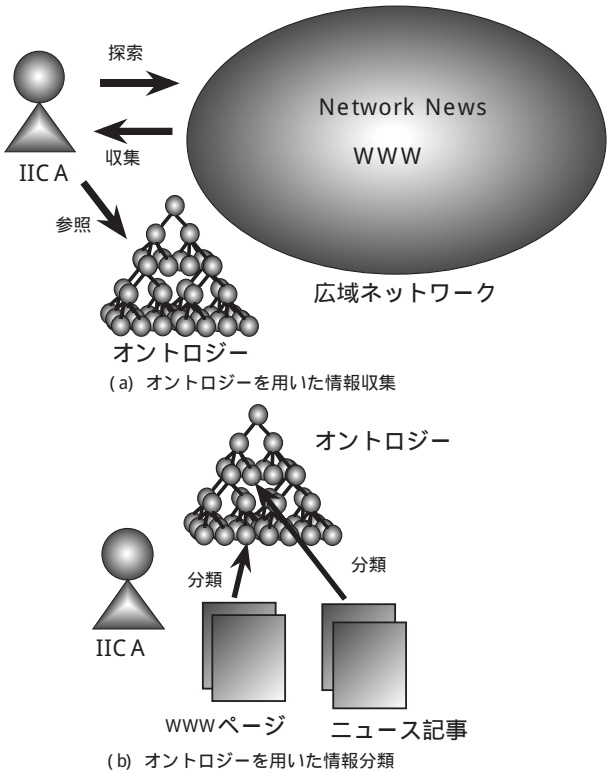

Fig 2: IICA:Intelligent Information Collector and Analyzer

2　弱構造化オントロジー

本章では, 情報収集・分類ために利用するオントロジーについて述べる.

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. Fig 3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-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.

弱構造化オントロジーを情報の収集と分類に用いることのメリットとしては, 次の 4 点がある.

1. 矛盾を含んだ情報の記述や処理を許容する頑健性
2. 知識メディア化によって人間にも機械にも容易に理解可能
3. 自然情報源からのボトムアップな構築の可能性
4. ユーザへのカスタマイズの可能性

本研究では, 約 4,500 の情報科学の専門用語から上記

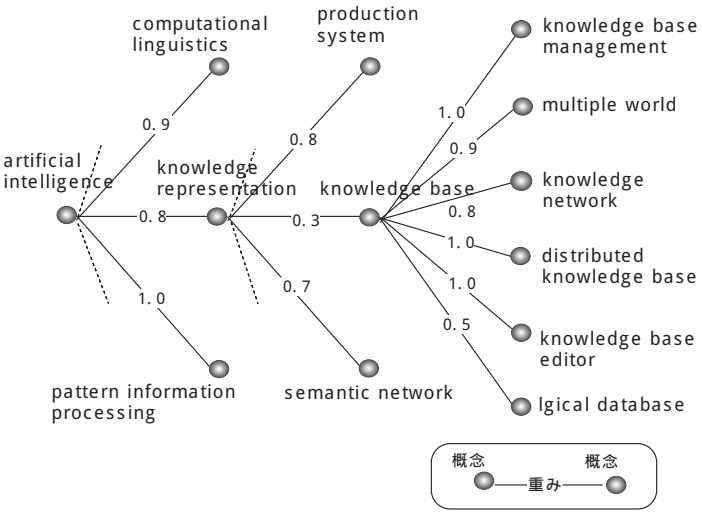

Fig 3: 弱構造化オントロジー

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　オントロジーを用いた情報収集

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. Fig 4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.

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
: 求める情報に関するキーワード列, 探索開始点となる

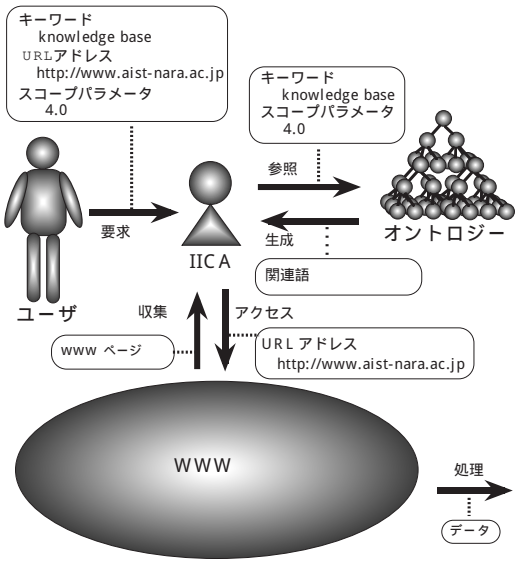

Fig 4: WWW における情報収集の概要

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.

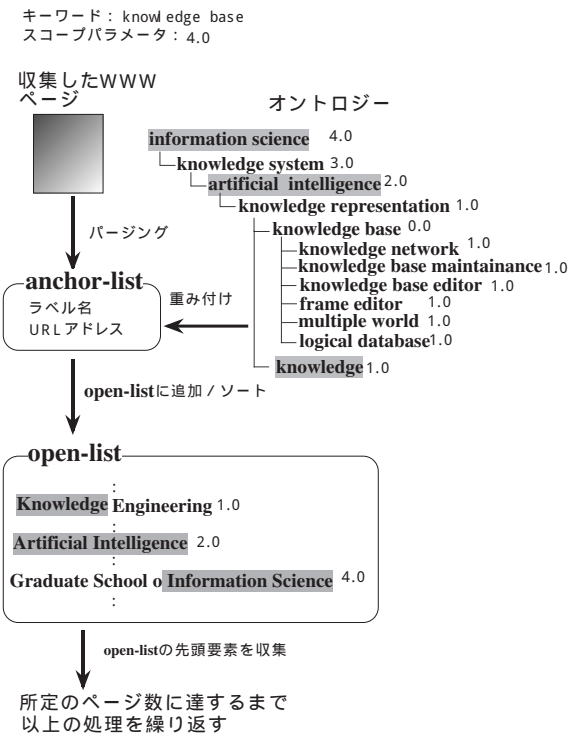

Fig 5: WWW における情報収集の例

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, HSFP 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.

具体例を用いて, 上記のアルゴリズムを説明する. いま, 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 5参照). IICA はまず, Fig 5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.

3.2 情報収集実験

上記の方法の妥当性を確認するために試験的な実験を試みた. “knowledge base” をキーワードに選び, ワークステション DEC 3000 alpha AXP を使用して, www を対象に IICA を実行した結果, , 約 12 時間で 500 ページを収集した. Table 1は、生成された関連語、その関連語を含んでいるページ数（割合）および入力キーワードからの距離（スコープ）を示している. この例では、収集した 90.2 % のページに入力キーワード “knowledge base” が含まれていたことがわかる. また、 Table 2は、入力キーワードとして “production system” と “semantic network” を与えた場合の結果を示す. この例では、収集したページの 56.0% および 17.6% に “production system” と “semantic network” が含まれていた. この結果は、前者の例に比べてページ数が少ない

が、 “production system” に関しては、オントロジーにおける “production system” の下位語である “rule base” および “production system” を比較的多くのページに含まれていることがわかる. 従来のキーワードサーチでは、入力したキーワードのみで収集を行なうため十分な情報が得られない可能性があるが、本手法では、オントロジー内の関連語も合わせて調べるため検索洩れが少ない. 本実験で収集したページの分類評価実験については、 4.3 節で述べる.

Table 1: www における情報収集実験の結果 - 1 キーワード

関連語	ページ数 (割合)	スコープ
knowledge	461 (92.2 %)	2.0
knowledge base	451 (90.2 %)	0.0
artificial intelligence	425 (85.0 %)	2.0
production system	252 (50.4 %)	2.0
knowledge representation	101 (20.2 %)	1.0
distributed knowledge base	90 (18.0 %)	1.0
blackboard model	72 (14.4 %)	2.0
semantic network	40 (8.0 %)	2.0
multiple world	26 (5.2 %)	1.0
knowledge network	11 (2.2 %)	1.0
predicate logic	10 (2.0 %)	2.0
frame	10 (2.0 %)	2.0
knowledge base management	2 (0.4 %)	1.0
knowledge base editor	0 (0.0 %)	1.0
logical database	0 (0.0 %)	1.0
inference control	0 (0.0 %)	2.0
representation of ambiguity	0 (0.0 %)	2.0

4 オントロジーによるテキスト分類

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.

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.

Table 2: www における情報収集実験の結果 - 2 キーワード

関連語	ページ数 (割合)	スコープ
artificial intelligence	382 (76.4 %)	2.0
knowledge	331 (66.2 %)	2.0
production system	280 (56.0 %)	0.0
rule base	178 (35.6 %)	1.0
production rule	140 (28.0 %)	1.0
knowledge base	123 (24.6 %)	2.0
semantic network	88 (17.6 %)	0.0
knowledge representation	54 (10.8 %)	1.0
working memory	30 (6.0 %)	1.0
blackboard model	27 (5.4 %)	2.0
IS-A link	20 (4.0 %)	1.0
PART-OF link	13 (2.6 %)	1.0
predicate logic	10 (2.0 %)	2.0
frame	10 (2.0 %)	2.0
spreading activation model	6 (1.2 %)	1.0
partitioned semantic network	4 (0.8 %)	1.0
inference control	2 (0.4 %)	2.0
representation of ambiguity	0 (0.0 %)	2.0

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, 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ている, , カ

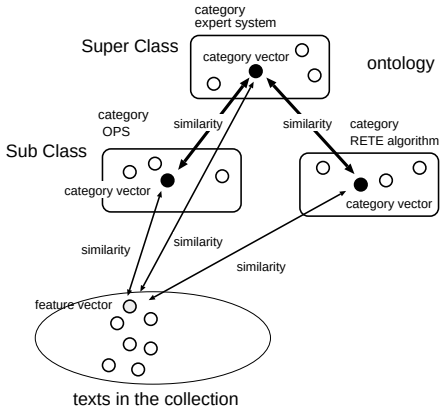

Fig 6: オントロジーによるテキスト分類

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 (Fig 6参照). 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. Table 3にその結果を示す. Table の右側は上位 20 カテゴリーの記事の数, 左の Table は下位 20 カテゴリーの記事の数である.

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.

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

Table 3: 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	7	chess machine	1
search	6	brain	1
fuzzy	6	bag	1
IEEE	6	backpropagation	1
backtracking	6	analog computer	1

Table 4: 分類実験の評価

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

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

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

計算結果を Table 4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、 Fig 8(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, 2 つ目グループは、 Fig 8(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 Table 5に示す.

最初の 2 つのグループに関しては, Fig 7に示すようなオ

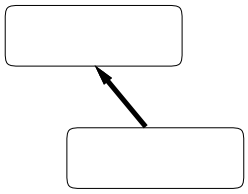

Fig 8: 誤り分類のグループ

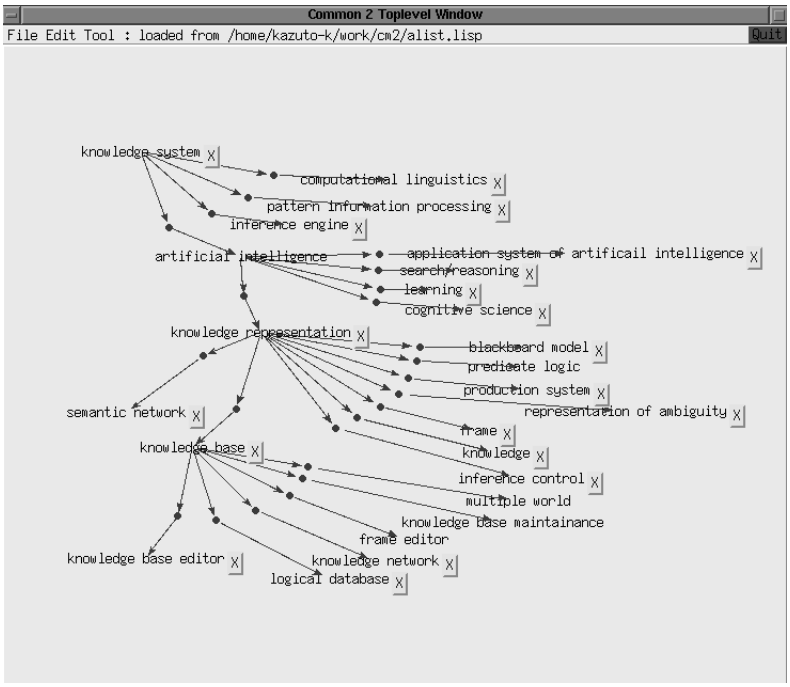

Fig 7: オントロジーブラウザ

Table 5: 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

Table 6: 分類実験の再評価

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

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, Table 6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.

また, 我々は, 3.3 の実験収集した 500 個の www 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

Table 7: 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

Table 8: www ページ分類実験の評価

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

その評価を, Table 7および Table 8に示す. 分類対象が, ドメインをかなり限定した入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, 本手法が, 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 つ利点があることが明らかになった.

- 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.
- エージェントはオントロジーを用いることでユーザの要求と関係のある情報を理解できる.
- 従来の方法では, 誤って分類された情報にアクセスすることはできなかったのに対し, 我々の方法では, オントロジーの階層関係をたどることで, 検索が可能になる.
- 弱構造化オントロジーは構築は, 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.

オンロジーを用いる本アプローチの有効性を利用者の専門性から分類すると Table 9になる. オントロジーと同じドメインを専門分野とする研究者や初学者はとっては、自動的収集した情報をオントロジーによって体系的に分類することは、情報の理解の助けとなり、有効であると言える. 一方、その分野に全くナイーブな人にとっては、オントロジーよる分類だけでは十分な理解の助けとはならない. この場合は、収集した情報と関連のある専門書や辞典とともに提示してやる必要があると思われる.

現在のシステムの問題としては, (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.

Table 9: 利用者の専門性と本システムの有効性

有効性	利用者のタイプ
○	専門家
○	初学者
△	非専門家

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. *Intoroduction to Modern Infomation 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.