# 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

# Ontology-Based Approach to Information Gathering and Text Categorization

岩爪道昭　　武田英明　　西田豊明

Michiaki Iwazume Hideaki Takeda Toyoaki Nishida

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

In this paper, we propose a new method of information gathering and text categorization using ontologies.

The number and diversity of information sources 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 have implemented a system called IICA (Intelligent Information Collector and Analyzer), which helps people to aquire knowledge from the information resources on the wide-area network by gathering information and categorizing texts. Evaluation of IICA is in progress. We have tested IICA for information gathering on the WWW and text categorization on the network news. We describe the result of the two tests.

## 1　はじめに

インターネットに代表される広域情報ネットワークの利用者の急速な増加はネットワーク上で提供される情報源の複雑化, 多様化, 大規模化もたらした. それにともない, 個人で処理しなければならない情報量も人間の能力を越えてきている. 我々が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. この問題に対処するために, さまざまな情報検索ツールの研究が行なわれ, 実際に利用可能になっている. 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.

そこで, 本研究では, 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. このシステムは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. さらに, 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. 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, 本アプローチが, 情報の収集と分類に有効であることが明らかになった.

## 2　オントロジーに基づくアプローチ

オントロジーは, 概念化の仕様を記述したものである [1]. 大規模知識共有のアプローチにおけるオントロジーの役割は, 人間とエージェントが知識の交換と情報のインデックス化を行なうための共通語彙を提供することである. 本研究では, 情報の収集と分類ためにオントロジーを利用する. したがってオントロジーの役割も次のように拡張される.

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の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, 既存の語彙体系をオントロジー構築の核として用いている. このようなオントロジーの大きな特徴は, 弱構造化 [6] すなわちクラス-インスタンス, 上位-下位, 部分-全体のような概念の関係や概念そのもの定義を持ないことである.

弱構造化オントロジーの記述例を図 1 に示す. これは, 人工知能に関するオントロジーの一部である. 最初の式を例に, 記述方法について説明する. オントロジーは, s 式で記述されている. con4001 はオブジェクトの内部 ID を表している. ( representation (symbol ‘‘knowledge base’’)) は, con4001 の外部表現がシンボル “knowledge base” であることを意味している. context は下位クラス・オブジェクトのリストを引数に持つ. 一つの要素は, 内部 ID と下位クラス・オブジェクトへの weight の値の組のリストになっている. weight は概念間のリンクの重み付けを与えるものである.

このように弱構造化オントロジーでは, 概念間の連想関係のみが記述されていため, オントロジーの構造を木からネッ

連絡先: 奈良先端科学技術大学院大学 情報科学研究科, 〒 630-01 生駒市高山町 8916-5, 電話:07437-9-9211 内線 5316, fax:07437-2-5269,

E-mail: mitiak-i@is.aist-nara.ac.jp, URL: http://ai-www.aist-nara.ac.jp/~mitiak-i/

```
(con4001 (representation (symbol ‘‘knowledge representation”))
        (context ((con4002 (weight 0.75))
                  (con4003 (weight 0.80))
                  (con4004 (weight 0.30))
                        :          ))))

(con4004 (representation (symbol ‘‘knowledge base”)
        (context ((con4017 (weight 0.80))
                  (con4018 (weight 1.00))
                        :
                  (con4022 (weight 0.50))))))

(con4017 (representation (symbol ‘‘knowledge network”)))
(con4018 (representation (symbol ‘‘knowledge base management”)))
                        :
(con4022 (representation (symbol ‘‘logical database”)))
```

図 1: 弱構造化オントロジーの記述例

トワークに拡張することが可能である.

弱構造化されたオントロジーを情報の収集と分類に用いるには次のようなメリットがある.

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 Intelligent Information Collector and Analyzer:IICA

我々は, 情報の収集とテキストの分類により, 広域ネットワークにおける情報資源からの知識獲得を支援する, IICA( Intelligent Information Collector andAnalyzer) と呼ぶシステムを実装した. この章では, IICA が, オントロジーを用いてどのように情報を収集し, 分類するか説明する. 評価実験によって本アプローチの有効性を示す.

### 3.1 オントロジーを用いた知的情報収集

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.

#### 3.1.1 www における情報探索

ここでは, IICA の www における情報探索の手順を示す. 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[1]の概要を示す.

[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 address, 文脈の推定範囲, 収集するページ数の入力.

[1] www のオフライン探索はネットワークに非常な負担をかけるので利用者は細心注意が必要である. 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なったり, 特定のホストに集中してアクセスしない, といったヒューリスティックが必要となる.

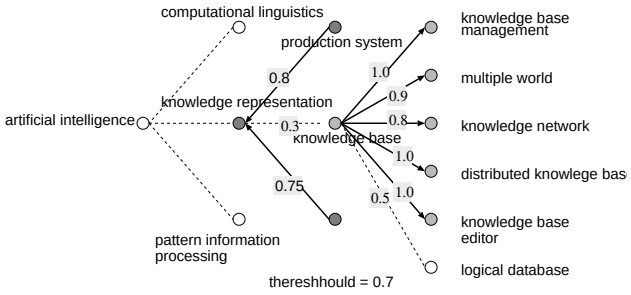


図 2: オントロジーを用いた知的な情報収集

[2] 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.

[3] 指定された url address 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. 存在する場合にはアーカイブから収集する. 存在しない場合は, HTTP にアクセスしてページを収集する.

[3’] 指定された url address を収集する.

[4] 収集したページが指定された数の満たなければ [5]. 達していれば探索終了.

[5] 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 address, ラベル名を抽出する. 各 url address 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.

[7] タイトルおよびラベルの中に [2] でリストアップされたキーワード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と数によってハイパーリンクに評価値を与える. ある一定の閾値を越えたものは open-list に加え評価値に従ってソーティングする.

[6] ハイパーリンクが収集ページにない場合には, open-list から url address を取り出す. [3’] へ.

#### 3.1.2 概念階層関係を利用した関連情報の推定

IICA が関連する情報をどのように推定するか図 2 に示す. ユーザからの入力がキーワード “knowledge base” であったする. もし, “knowledge base” という語彙を含むテキストが存在しなかった場合, IICA は入力されたキーワードと意味的に関連性を持つ語彙を推論し, “knowledge base management”, “distributed knowledge base”, “knowledge base editor” だけでなく, “knowledge network”, “multiple world” などに関するテキストを収集する.

IICA はまた, ユーザの質問から文脈の推定の行なう. 例えば, キーワード列入力が “semantic network” と “logical database” のとき, 文脈が “knowledge representation” であると解釈する.

文脈に関する推論は, 概念のレベル ( オントロジー内での位置 ) と概念間のウエイトに依存する. ユーザはしきい値を設定することで, 文脈の推定範囲を制御することができる.

#### 3.1.3 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. 分析の結果, 収集した 90.2 % のページが “knowledge base” を含み, 本方法が妥当であることが明らかになった. 現在, さらにいろいろなケースの入力よる評価実験を進めている.

### 3.2 オントロジーによるテキスト分類

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,

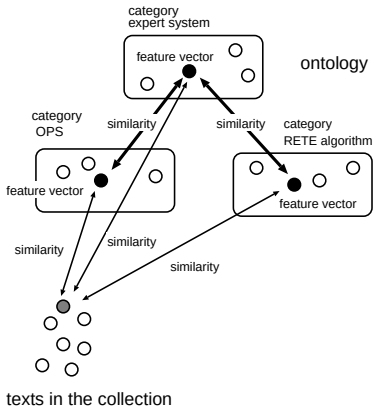


図 3: オントロジーによるテキスト分類

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. 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.

### 3.2.1 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, カテゴリー間の意味的な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 3 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.

[1] 収集したテキストの特徴ベクトルの計算.
[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なるキーワードマッチングで分類する.
[3] 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.
[4] 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.
[5] 各カテゴリの特徴ベクトルが収束するまで [3],[4] を繰り返す.
[6] 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.

### 3.2.2 ベクトル空間モデル

上記の方法で用いている特徴ベクトルの計算には, IDF ( Inverse Document Frequnecy )[4] 使用している. この方法は, 多くのドキュメントに出現する単語の重要度をさげ, より少ないドキュメントに現れる単語を強める特徴がある. キーワードの重みは $w_{ik}$ は次の式で与えられる.

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

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.

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

表 1: 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 |

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.

### 3.2.3 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 1 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.

表 2: 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, Accuracy(A), Recall(R) および Precision(P) を次の式を用いて計算した,

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

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

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

表 3: 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 |

計算結果を表 2 に示す, ただし, Recall と Precision が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 つのケースに大別した. 最初のケースは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ケース, 2 つ目ケースは正しく分類されべきクラスの上位クラスに分類されたケース, 最後は, 正しく分類されるべきクラス

表 4: 表 3 の結果を考慮に入れた評価

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

とは全く無関係のクラスに分類されたケースである．その分析結果を表 3 に示す．

最初の 2 つのケースは、オントロジーにおける意味的関係を辿ることで、誤って分類された情報にもアクセス可能であることを示すものである．そこで、この 2 つのケースを正解と仮定して再評価した結果、表 4 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．

## 4 議論と展望

### 4.1 オントロジーと知識メディア

我々は, ヒューマンメディアとコンピュータメディアを統合する知識メディアの利用を試みている．図 3 は前節の記事分類の実験結果の一部を可視化したものである．図中の小さな円は, 記事 1 つを意味している．各円から出ている矢印は各記事の内容を示すキーワードを指している．結果の可視化は, 記事どうしのオントロジカルな関係の理解を可能にする．IICA は知識メディアとして実現されたオントロジカルな検索ツールを持っており, 非専門家のユーザでも情報の検索, 分類結果の理解が容易になる．

### 4.2 未知語処理とオントロジーのカスタマイズ

表 5: 未知語の出現頻度

| signifacant words in the unknown words list | |
|---|---|
| ontology | 27 |
| artificial life | 9 |
| Mosaic | 7 |
| internet | 8 |
| Lucid | 5 |
| KIF | 5 |
| WWW | 5 |

本アプローチではテキスト中の出現頻度が高く, キーワードの候補になった場合でも, オントロジーに存在しない場合には、未知語として処理される．このようにして生成される未知語のリストを分析した結果, オントロジーにない重要な概念や比較的新しい概念が含まれていることがわかった (表 5 参照)．現在, この未知語からの知識獲得をどこまで自動化できるか検討している．

また, 収集したデータから, 概念間の重みの変更や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についても検討している．

## 5 まとめ

本研究では, オントロジーを用いた, 新しい情報の収集とテキスト分類方法を提案した．情報を収集し, テキストを分類することによって広域情報ネットワークからの知識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 IICA 用いた www における情報収集とネットワークニュースにおける分類実験を行ない, 本方法が有効であることを確認した．今後は, オントロジーのカスタマイズ, 未知語からの新出語の発見などに取り組む予定である．

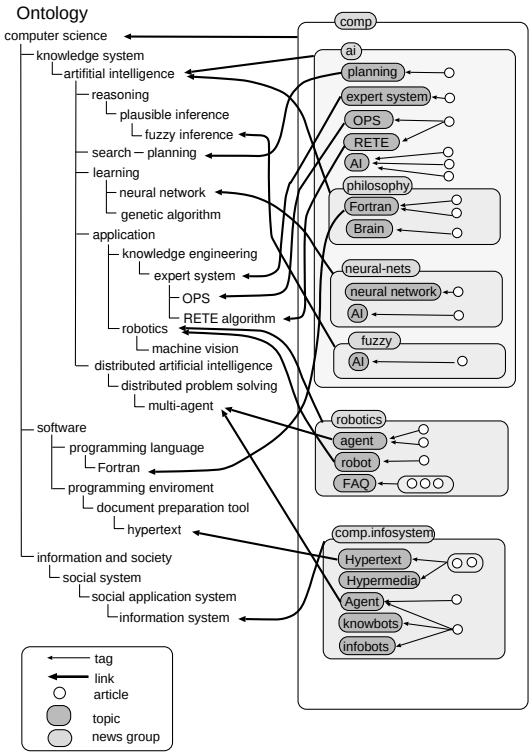


図 4: オントロジーの役割と知識メディア

## 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.